Model-Master-Transfer: Formally Deconstructing Educational Games to Build a Quantitative Theory

Dominicus Tornqvist

BA Hons

School of Information and Communication Technology
Griffith Sciences
Griffith University

Submitted in fulfilment of the requirements of the degree of Doctor of Philosophy
April 19, 2020
Abstract

We live in an increasingly interconnected, complex world, where our collective decisions can have unanticipated indirect consequences on the world. What is today a hands-on job may soon be a job of managing variables at a computer screen. However, people have tremendous difficulty handling even relatively mild levels of complexity in experiments. Yet, in the context of many modern computer games, people eagerly teach themselves vastly complex systems, all without any external guidance or coercion. Educational video games research is a menagerie of different methodologies and paradigms. Meta analyses have found that mixed results in the literature are difficult to interpret due to the combination of different theoretical approaches, different data reporting conventions, and a general focus on proof-of-concept studies. They recommend a transition to narrower investigations of specific causal relationships between game properties and outcomes of engagement and learning, but this requires a way to incorporate them into the broader picture of educational and serious games research.

In this thesis, I focus on educational games and propose the Model-Master-Transfer (MMT) framework to break down educational game usage into a set of formal subprocesses that can be studied in more depth individually, and specify how such narrow studies can then be assembled to build up a causal model of the underlying effects. The framework is illustrated using different educational examples. The conceptual study contributes a comprehensive framework to the ambiguous research on educational learning using games. MMT is then used in empirical experiments to address two sub-problems: 1) Why players sometimes choose to lose the game, completely derailing its intended purpose; 2) The design of inherently learnable systems in terms of how the complexity of a game relates to the player’s ability to master it - searching for forms of complexity that elicit curiosity to learn about that complexity. These experiments demonstrate the value and manner of how to apply MMT to investigate specific psychological phenomena while retaining a logically coherent place in our understanding of how educational and serious games achieve positive outcomes - a logically coherent place provided by the framework of MMT. This body
of work provides practical applications for game developers and educators alike, as well as interesting theoretical implications for cognition, curiosity, and complexity.
Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Dominicus Tornqvist

August 7, 2020
Contents

Declaration i

List of Figures viii

List of Tables xi

List of Abbreviations xiii

List of Publications xv

1 Introduction 1
  1.1 Problems ................................................................. 1
  1.2 Research Questions .................................................. 3
  1.3 Aims ........................................................................... 4
  1.4 Approach ................................................................. 6
  1.5 Contributions ........................................................... 13
  1.6 Structure ................................................................. 17

2 Literature Review 19
  2.1 Learning ................................................................. 20
     2.1.1 System Dynamics & Complex Problem Solving ............. 20
     2.1.2 Education via Games ............................................. 22
     2.1.3 Implicit vs. Explicit Theories of Learning for Games ...... 23
     2.1.4 Ambiguity of Educational Game Outcomes ................. 27
     2.1.5 Entities & Processes in Existing Frameworks ................ 31
  2.2 Game Design ........................................................... 39
     2.2.1 Games vs. Simulations ........................................... 39
3  Proposing the Model-Master-Transfer Framework 70
3.1 Defining Model-Master-Transfer 71
3.2 Mathematical Structure Overview 73
3.3 Usage of MMT: Measurements & Predictions 78
  3.3.1 Modelling 78
  3.3.2 Mastery 84
  3.3.3 Transfer 84
  3.3.4 Engagement & Play Appeal 87
  3.3.5 Learning Difficulty & Complexity 90
3.4 Implications & Applications 93
  3.4.1 Implications for Practitioners 93
  3.4.2 Implications for Researchers 96
3.5 Limitations & Future Work 100
  3.5.1 Modelling 101
  3.5.2 Mastery Through Play 101
  3.5.3 Transfer 102
  3.5.4 Social Context 102
3.6 Conclusion 103

4  Proposing Dynamic Causal Nets: Calculating Game Kolmogorov Complexity 106
4.1 Introduction 106
4.2 Proposing Dynamic Causal Nets 107
4.3 Calculating Kolmogorov Complexity 111
4.4 Conclusion 116
5 Study 1: Evaluating Play Appeal Factors & Engagement
5.1 Introduction: Motivated to Lose? ........................................... 118
5.2 Literature Review: Concepts of Motivation in Play .................. 121
  5.2.1 Game Value Adoption .................................................. 122
  5.2.2 Challenge ................................................................. 123
  5.2.3 Juice ........................................................................... 124
5.3 Study Framework & Definitions ............................................. 126
  5.3.1 Different Win-Lose Definitions ....................................... 126
  5.3.2 Defining Play Appeal Factors: Proposing Dynamic Probability
    Response ................................................................. 129
5.4 Method .............................................................................. 138
  5.4.1 Game ................................................................. 138
  5.4.2 Measure ................................................................. 139
  5.4.3 Checking the Game’s Challenge Curves ......................... 142
  5.4.4 Participants ............................................................ 146
  5.4.5 Procedure ............................................................... 146
5.5 Results .............................................................................. 148
5.6 Discussion ......................................................................... 154
  5.6.1 Play Motivations Application & Theory ......................... 158
  5.6.2 Limitations & Future Work ......................................... 160
5.7 Integrating With MMT ....................................................... 162
5.8 Conclusion ......................................................................... 163

6 Study 2: Game Complexity & Curiosity ................................ 166
6.1 Introduction ......................................................................... 167
6.2 Hypotheses From Theories ................................................. 170
  6.2.1 Emergent Gameplay & Epistemic Emergence .................. 170
  6.2.2 Play & Curiosity ........................................................ 176
  6.2.3 Exploratory Hypotheses .............................................. 177
6.3 Method: Experiment 2.1 ..................................................... 179
  6.3.1 Game Design ............................................................ 180
  6.3.2 Checking Levels of Static & Dynamic Complexity .......... 184
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.3</td>
<td>Measures</td>
<td>193</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Procedure</td>
<td>197</td>
</tr>
<tr>
<td>6.4</td>
<td>Results: Experiment 2.1</td>
<td>200</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Dynamic Complexity Game</td>
<td>200</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Static Complexity Game</td>
<td>201</td>
</tr>
<tr>
<td>6.5</td>
<td>Discussion of Experiment 2.1</td>
<td>202</td>
</tr>
<tr>
<td>6.6</td>
<td>Method: Experiment 2.2</td>
<td>205</td>
</tr>
<tr>
<td>6.6.1</td>
<td>Game Design</td>
<td>206</td>
</tr>
<tr>
<td>6.6.2</td>
<td>Measures</td>
<td>206</td>
</tr>
<tr>
<td>6.6.3</td>
<td>Procedure</td>
<td>208</td>
</tr>
<tr>
<td>6.6.4</td>
<td>Evaluating the Curiosity Measure CEI-II</td>
<td>209</td>
</tr>
<tr>
<td>6.7</td>
<td>Results: Experiment 2.2</td>
<td>210</td>
</tr>
<tr>
<td>6.7.1</td>
<td>Dynamic Complexity Game</td>
<td>210</td>
</tr>
<tr>
<td>6.7.2</td>
<td>Static Complexity Game</td>
<td>212</td>
</tr>
<tr>
<td>6.8</td>
<td>Discussion of Experiment 2.2</td>
<td>215</td>
</tr>
<tr>
<td>6.9</td>
<td>Discussion</td>
<td>219</td>
</tr>
<tr>
<td>6.9.1</td>
<td>Practical Applications in Educational Game Design</td>
<td>219</td>
</tr>
<tr>
<td>6.9.2</td>
<td>Theoretical Implications</td>
<td>221</td>
</tr>
<tr>
<td>6.9.3</td>
<td>Limitations &amp; Future Work</td>
<td>232</td>
</tr>
<tr>
<td>6.10</td>
<td>Integrating With MMT</td>
<td>235</td>
</tr>
<tr>
<td>6.11</td>
<td>Conclusion</td>
<td>236</td>
</tr>
<tr>
<td>7</td>
<td>Discussion</td>
<td>239</td>
</tr>
<tr>
<td>7.1</td>
<td>Primary Research Question</td>
<td>239</td>
</tr>
<tr>
<td>7.2</td>
<td>Dynamic Causal Nets &amp; Kolmogorov Complexity</td>
<td>244</td>
</tr>
<tr>
<td>7.3</td>
<td>Research Question 2: Motivated to Lose</td>
<td>244</td>
</tr>
<tr>
<td>7.4</td>
<td>Research Question 3: Complexity &amp; Curiosity</td>
<td>246</td>
</tr>
<tr>
<td>7.5</td>
<td>Practical Applications: Educational Game Design</td>
<td>249</td>
</tr>
<tr>
<td>7.5.1</td>
<td>Objectively Quantifiable Game Metrics for Game Design</td>
<td>249</td>
</tr>
<tr>
<td>7.5.2</td>
<td>Play Motivation Interactions</td>
<td>250</td>
</tr>
<tr>
<td>7.5.3</td>
<td>Relationship Between Complexity, Learning &amp; Engagement</td>
<td>250</td>
</tr>
<tr>
<td>7.5.4</td>
<td>Curiosity as a Play Motivation</td>
<td>250</td>
</tr>
</tbody>
</table>
7.6 Theoretical Implications .............................................. 251
  7.6.1 Conceptions of Complexity in Games ......................... 251
  7.6.2 Emergent Games: Inherently Learnable Systems? .......... 252
  7.6.3 Are Players Curious About Complexity? ...................... 252
  7.6.4 Curiosity & Play Motivation Interactions ................... 252
7.7 Methodological Implications ...................................... 253
7.8 Limitations & Future Work ......................................... 254
  7.8.1 Primary Research Question: MMT ......................... 254
  7.8.2 Research Question 2: Motivated to Lose .................... 256
  7.8.3 Research Question 3: Complexity & Curiosity .............. 258

8 Conclusion ..................................................................... 259

A Massively Scalable Learning ........................................ 264
  A.1 Introduction ......................................................... 265
  A.2 Background: The Problem ....................................... 266
    A.2.1 Individual Differences & Individual Learning Efficiency .... 266
    A.2.2 Unguided Exploration in Education ....................... 268
  A.3 What is Massively Scalable Learning? ......................... 268
    A.3.1 Applications & Uses ......................................... 270
    A.3.2 Factors Determining MSL Effectiveness ................... 273
  A.4 Implementation: Design Principles & Guidelines ............ 275
    A.4.1 Removing Obstacles & Minimising Contingencies .......... 275
  A.5 Components of MSL .............................................. 281
    A.5.1 Learnable ..................................................... 282
    A.5.2 Practical ..................................................... 283
    A.5.3 Motivational .................................................. 285
  A.6 Conclusion .......................................................... 287

B Cognitive Habitat, Strategy Ecosystem: A Predictive Model ... 289
  B.1 Introduction ........................................................ 290
  B.2 Literature Review .................................................. 291
  B.3 CHSE Model ....................................................... 292
# List of Figures

1.1 Educational Game Research ........................................... 8
1.2 Quantifying Game Properties ........................................... 9
1.3 The Focus ............................................................... 10
1.4 The Approach ............................................................ 12
1.5 Thesis Structure Overview .............................................. 18

2.1 Simple CBN Example ..................................................... 25
2.2 Educational Game Literature ............................................ 30
2.3 U-Shaped Relationship of Challenge .................................. 49
2.4 U-Shaped Relationship of Chaos ...................................... 65
2.5 U-Shaped Relationship of Network Connectivity ................... 67

3.1 Model-Master-Transfer Framework ................................... 72
3.2 Causal Model Comparison Example .................................. 83
3.3 MMT Formula Summary ................................................. 92

4.1 Simple Example Game Noun-Verb Network ......................... 108
4.2 DCN Simple Example .................................................... 110
4.3 DCN Aviation Example .................................................. 111

5.1 2D Difficulty Space ....................................................... 131
5.2 Example Challenge Curves .............................................. 133
5.3 Game Screenshot .......................................................... 139
5.4 Relative Player Preference Calculation ............................... 140
5.5 Measured 2DCC ............................................................ 143
5.6 Measured 2DDS ............................................................ 144
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.8</td>
<td>Example Strategy Ecosystem Graph</td>
<td>311</td>
</tr>
<tr>
<td>B.9</td>
<td>Strategy Transition Network</td>
<td>324</td>
</tr>
<tr>
<td>B.10</td>
<td>Area of Applicability Affects Utility</td>
<td>330</td>
</tr>
<tr>
<td>B.11</td>
<td>Example Situation Frequency Graph</td>
<td>331</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Number</th>
<th>Table Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Summary of Reviewed Frameworks</td>
<td>37</td>
</tr>
<tr>
<td>5.1</td>
<td>Study 1 Measure</td>
<td>141</td>
</tr>
<tr>
<td>5.2</td>
<td>Study 1 Descriptive Statistics</td>
<td>148</td>
</tr>
<tr>
<td>5.3</td>
<td>Study 1 Planned Comparisons</td>
<td>151</td>
</tr>
<tr>
<td>5.4</td>
<td>Study 1 Results Summary</td>
<td>154</td>
</tr>
<tr>
<td>6.1</td>
<td>Experiment 2.1 Conditions</td>
<td>179</td>
</tr>
<tr>
<td>6.2</td>
<td>Lyapunov Exponent Measurements</td>
<td>190</td>
</tr>
<tr>
<td>6.3</td>
<td>Experiment 2.1 Measure Definitions</td>
<td>196</td>
</tr>
<tr>
<td>6.4</td>
<td>Experiment 2.1 Demographics</td>
<td>198</td>
</tr>
<tr>
<td>6.5</td>
<td>Experiment 2.1 Dynamic Complexity Pairwise Comparisons</td>
<td>200</td>
</tr>
<tr>
<td>6.6</td>
<td>Experiment 2.1 Static Complexity Pairwise Comparisons</td>
<td>201</td>
</tr>
<tr>
<td>6.7</td>
<td>Experiment 2.1 Descriptive Statistics</td>
<td>202</td>
</tr>
<tr>
<td>6.8</td>
<td>Experiment 2.1 Hypotheses &amp; Results</td>
<td>203</td>
</tr>
<tr>
<td>6.9</td>
<td>Experiment 2.2 Conditions</td>
<td>206</td>
</tr>
<tr>
<td>6.10</td>
<td>Experiment 2.2 Measure Definitions</td>
<td>207</td>
</tr>
<tr>
<td>6.11</td>
<td>Experiment 2.2 Demographics</td>
<td>208</td>
</tr>
<tr>
<td>6.12</td>
<td>CEI-MkII Item Weightings</td>
<td>210</td>
</tr>
<tr>
<td>6.13</td>
<td>Experiment 2.2 Static Complexity Pairwise Comparisons</td>
<td>213</td>
</tr>
<tr>
<td>6.14</td>
<td>Experiment 2.2 Descriptive Statistics</td>
<td>214</td>
</tr>
<tr>
<td>6.15</td>
<td>Experiment 2.2 Hypotheses &amp; Results</td>
<td>215</td>
</tr>
<tr>
<td>6.16</td>
<td>Study 2 Hypotheses &amp; Results Summary</td>
<td>218</td>
</tr>
<tr>
<td>6.17</td>
<td>Study 2 Follow-up Conditions &amp; Hypotheses</td>
<td>234</td>
</tr>
<tr>
<td>A.1</td>
<td>Major Components of MSL</td>
<td>281</td>
</tr>
</tbody>
</table>
B.1 Strategy Monitoring Costs .................................................. 327
B.2 Strategy Switching Costs .................................................. 328
List of Abbreviations

2DD: Two-Dimensional Difficulty (see page 130)
2DDS: Two-Dimensional Difficulty Space (see page 130)
ACS: Academic Curiosity Scale (see page 44)
AI: Artificial Intelligence (see page 142)
ANOVA: Analysis of Variance (see page 152)
ATMSG: Activity Theory-based Model of Serious Games (see page 37)
CBGD: Cognitive Behavioural Game Design (see page 37)
CBN: Causal Bayes Net (see page 24)
CEI-II: Curiosity Exploration Inventory Mark II (see page 44)
CHSE: Cognitive-Habitat Strategy-Ecosystem (see page 289)
CMQ: Current Mood Questionnaire (see page 43)
CPS: Complex Problem-Solving (see page 21)
DCN: Dynamic Causal Nets (see page 107)
DES: Differential Emotions Scale (see page 43)
DPR: Dynamic Probability Response (see page 129)
DRM: Digital Rights Management (see page 283)
EEG: Electroencephalography (see page 43)
ENED-GEM: Environmental Educational Game Enjoyment Model (see page 37)
GEngQ: Game Engagement Questionnaire (see page 45)
GExpQ: Game Experience Questionnaire (see page 45)
GOM II: Game Object Model version II (see page 37)
GSD: Global Software Development (see page 113)
GSR: Galvanic Skin Response (see page 43)
IBLT: Instance-Based Learning Theory (see page 26)
IMI: Intrinsic Motivation Inventory (see page 44)
KC: Kolmogorov Complexity (see page 112)
LM-GM: Learning Mechanics- Game Mechanics (see page 37)
MAACL: Mood and Affect Checklist (see page 43)
MAACL-R: Mood and Affect Checklist, Revised version (see page 43)
MANCOVA: Multivariate analysis of covariance (see page 210)
MMT: Model-Master-Transfer (see page 71)
MSL: Massively Scalable Learning (see page 264)
NA: Not Applicable (see page 198)
NES: Novelty Experiencing Scale (see page 44)
PANAS: Positive and Negative Affect Schedule (see page 43)
PANAS-X: Positive and Negative Affect Schedule, Expanded (see page 44)
PC: Personal Computer (see page 146)
PENS: Player Experience of Need Satisfaction (see page 45)
POMS: Profile of Mood States (see page 43)
RCT: Randomised Controlled Trial (see page 11)
RETAIN: Relevance Engagement Translation Assimilation Immersion Naturalisation (see page 37)
SD: Standard Deviation (see page 148)
SECS: State Epistemic Curiosity Scale (see page 44)
SEG: Strategy Ecosystem Graph (see page 311)
STPI: State-Trait Personality Inventory (see page 44)
SSS: Sensation Seeking Scale (see page 44)
TETEM: Technology-Enhanced Training Effectiveness Model (see page 37)
TWC: Total Walk Count (see page 173)
List of Publications

This research has resulted in the following publications:


  – Abstract: Until recently, research on video games has been concerned largely with negative effects of play. Increasingly, however, a range of positive psychological effects of playing video games are being reported. Being able to identify positive impacts is an important step toward leveraging the huge appeal of game playing to aid psychological well-being. This chapter discusses a range of benefits associated with playing two genres of popular off-the-shelf video games as opposed to video games that have been constructed to specifically target education or teach health-related lessons. With the video game phenomenon set to have an ever-increasing impact across society on a global scale, knowledge of games that have the potential to meet positive psychological needs will be critical to leveraging positive outcomes from video game play.


  – Abstract: An application for which serious games are almost uniquely suited, and yet rarely mentioned, is Massively Scalable Learning (MSL):
the design of an educational object that is effective at delivering its value in virtually any conditions, minimizing any dependencies or requirements such as external guidance. This allows it to be shared and spread, maximizing its reach to the human population. Games (particularly exploratory games) can be well suited to this educational goal. The reasons for this are discussed in this paper, along with the advantages of designing inherently learnable systems, which domains of knowledge would benefit from MSL, and some of the design implications in making games for MSL.


  – Abstract: How can one measure the learning outcome of playing a serious game? We need an objective measure of the learning contents included in a game. The research is diverse, utilizing vastly different games to teach various different kinds of knowledge and skills. This makes it difficult to compare and generalize studies, lacking any established formal tool of analysis. This problem requires the design of an abstract and objective measurement of the quantity of learning material independent of the learning domain. Based on cognitive research on causal Bayes nets (CBNs), this paper proposes using dynamic causal nets (DCNs) to model an abstract knowledge base, which could be mapped to many different domains of learning. We also apply Kolmogorov Complexity (KC) as an approach to measure the content of the abstract knowledge base. This work will establish a theoretical foundation for future research of serious games.


  – Abstract: Players report losing some games can be as much fun, or more than winning. It is imperative to identify what motivates a player to pur-
sue failure in games due to the importance of many games now used for educational and health purposes. The game’s intended outcomes can be entirely undermined if players would rather lose than win the game. To achieve reliable predictions on the win/lose dilemma, we propose a new model of challenge, Dynamic Probability Response, which quantifies the degree and type of challenge. Many previous studies focus on individual differences in play. This study focuses on how different play motivations interact. Three conceptualisations of winning were tested against each other by giving players a mutually exclusive choice between challenge, gratuitous feedback from interaction (juice hypothesis), and compliance with visual cues denoting victory (game value adoption hypothesis). Each potential motivation for play was derived from psychological theory that is prevalent in the game design literature. Using a within-subject ANOVA, the three hypothesised motivations investigated were each individually supported. Some hypotheses about which motivations can disrupt the game’s goal were supported. Others were not. The applications of these results to game and simulation design are discussed.

• Tornqvist, Dominicus, Wen, Lian, & Tichon, Jennifer. In review. "Massively Scalable Learning: Priorities and Principles of Serious Game Scalability."

Abstract: There is a healthy research community focused on individual differences to tailor serious games for maximum effect for each person. But there is a comparative lack of research on the scalability of serious games for maximising knowledge saturation in a population. Scalability is critical in many real applications. The authors detail this neglected set of priorities as a research paradigm: Massively Scalable Learning (MSL), and delineate what kinds of domains would benefit most from MSL, summarise its specific cognitive, motivational, and practical components, and detail the factors and mechanisms that determine if MSL is relevant and effective. Existing research is examined, evaluating common educational tools and game features such as virtual tutors for their applicability to MSL, to extract some initial guidelines and principles for MSL for prac-
titioners of serious game design, and identify the key knowledge gaps where future research needs to focus in this neglected but important area.


  Abstract: Educational video games research is a menagerie of different methodologies and paradigms. Mixed results in the literature are difficult to interpret due to the combination of different theoretical frames, different data reporting conventions, and a general focus on proof-of-concept studies. Narrower investigations of specific causal relationships between game properties and outcomes of engagement and learning can provide more specific findings, but require a way to incorporate them into the broader picture of educational video games research. In this study, we focus on educational video games and propose the Model-Master-Transfer (MMT) framework to break down educational video game usage into a set of formal sub-processes that can be studied in more depth individually, and specify how such narrow studies can then be assembled to build up a causal model of the underlying effects. The framework is illustrated using different educational examples. The conceptual study contributes a comprehensive framework to the ambiguous research on educational learning using video games. Additionally, it provides practical implications for game developers and educators alike.


  Abstract: Cognitive research has found people are sometimes adept and sometimes inept at handling complexity. Complexity is a key concept in much of cognitive science, yet the field has scarcely incorporated any of the work in complexity theory. Complexity theory may generally be too abstract to easily apply to human cognition studies. Here, the problem is addressed by considering complexity in terms of the related con-
cept of emergence, building a model of epistemic emergence, Cognitive-Habitat Strategy-Ecosystem, (CHSE) to act as an overarching framework into which different conceptions of complexity and cognition can be integrated, describing how they will then interact to produce implications and predictions for cognition in complex systems. CHSE is used to derive a definition of an emergent system based on how it would be experienced as complex. This model provides value both at the micro level, by generating specific predictions, and at the macro level, through hypothesising interactions between other cognitive theories such as cognitive load, and adaptation from failure. We detail the model’s assumptions, functionality, and possible ways to measure relevant variables.


Abstract: The design of effective educational games would be greatly enhanced by a scientific understanding of how curiosity affects interest and engagement with games with different types of complexity. This study investigated the relationship between game system complexity and engagement, given that many people teach themselves complex games for fun, and yet have trouble understanding complex systems in the lab. What is the relationship between game system complexity and engagement, and how is it affected by the curiosity of the learner? Gathering theories from psychology, complexity theory, and game design produces conflicting answers. Definitions of emergence suggest an interesting model whereby curiosity peaks at a medium level of complexity (Emergent Gameplay hypothesis, from theories of curiosity as incongruity and optimal arousal). But other theories of play and curiosity suggest that players will prefer high levels of complexity (Challenging Complexity hypothesis, from drive theories of curiosity), or prefer lower levels of complexity (Effectance hypothesis, from theories of play as effectance). These conflicting theories were tested with a within-subjects experimental design, using three different complexity levels (low, medium, and high).
for two different conceptions of complexity (dynamic and static), then collecting participants ratings of preference and interest, and their trait curiosity using the Curiosity Exploration Inventory Mark II (CEI-II). The Emergent Gameplay and Challenging Complexity hypotheses were not supported, failing to find support for the drive, optimal arousal, or incongruity theories of curiosity behind those hypotheses. But there was some support for the Effectance hypothesis, thereby supporting the effectance theory of play. Individuals’ trait curiosity showed significant interactions with measures of engagement, with more curious individuals generally more interested regardless of complexity level, which was not consistent with any theories of curiosity, but suggests that in practical applications, a focus on tailoring to less curious individuals could generate more value in terms of overall engagement with the material. Theoretical implications are discussed and practical applications in educational game design outlined. Work remains to develop theories to fully explain why and predict which different forms of complexity (e.g. static or dynamic) might produce different outcomes of engagement via curiosity.


  - Abstract: There is a kind of seemingly nonsensical play behaviour found in the simulation sandbox game genre. This behaviour is very spontaneous and impulsive and associated with self-initiated learning, and here the author seeks to better understand what it is, and why it seems associated with simulation sandbox games. That is: What purpose does it serve, and what might provoke it? This requires a review of the literature on this play behaviour, and that of the simulation sandbox genre, respectively. From this review the author concluded firstly that there is strong support that exploratory play is for discovering the structure and behaviour of systems, and secondly that the observable characteristics of exploratory
play make it a highly probable candidate for the bizarre behaviour observed in simulation sandbox games. Moreover, several hypotheses were generated by identifying many characteristics of the genre (e.g. system complexity and responsiveness) that are directly relevant to the theorized motivations for exploratory play, suggesting some directions for future research into what conditions and designs might encourage exploratory play. Knowledge of the relationship between this genre and this form of play could prove invaluable for designing games for learning, because despite being centrally relevant to many studies on game-based learning, exploratory play has been neglected. It has been needlessly isolated in distinct strands of research on its components, which will be unified here to provide a comprehensive account of this behaviour and its importance to future research in this area.
Chapter 1

Introduction

Virtually every domain of modern life has been increasing in complexity. Citizens across the world have become entangled in global economic and environmental issues such as climate change, and it seems almost every profession is constantly adapting to new technologies that require experts learn new ways of doing things, such as remote robotic surgery techniques, or new aviation autopilot systems. What is now a hands-on job may soon be a job of managing variables at a computer screen. Additionally, citizens are now faced with the indirect, long-term consequences of their collective decisions on the environment. Investigating such complex issues in our interconnected world requires an interdisciplinary approach to research, which is why this thesis spans fields such as motivational psychology, complexity theory, and cognitive science. Increasingly, humans are faced with unfamiliar complex systems that they must master. Given the popular appeal of many complex games that players seem to master effortlessly, it is no surprise that increasing attention is focusing on using games for educational purposes. The field has now produced quite a diverse body of work, utilising various methodologies, measures, and theories.

1.1 Problems

There is much diversity in serious games research not just in terms of methodologies, but also in their outcomes, making the body of literature very difficult to in-
Multiple literature reviews and meta-analyses (Boyle et al., 2016; Clark, Tanner-smith, and Killingsworth, 2014; Connolly et al., 2012; McClarty et al., 2012) have noted this problem and recommend research shift away from proof-of-concept studies toward narrower studies of how a specific design change affects specific outcomes such as engagement and learning. Narrower studies of very specific phenomena do exist, but do not often have obvious application to the broader question of how serious games should be designed or used. Is there a way serious games can be broken down into narrower sub-processes for individual study, and at the same time improve the integration of such narrow studies to form a coherent larger picture?

Many games are staggeringly complex and yet players eagerly teach themselves to master them without any supervision. This self-motivating property is rightly seen as a powerful property of games. Although games have been used to teach a variety of subjects (e.g. science, Cheng and Su, 2012; Ronen and Eliahu, 2000; Squire et al., 2004, spatial skills, Green, Pouget, and Bavelier, 2010; Shute, Ventura, and Ke, 2015, persistence, Shute, Ventura, and Ke, 2015; Ventura, Shute, and Zhao, 2013, and metacognition Kim, Park, and Baek, 2009; Moncarz, 2011), there have also been plenty of negative and mixed results, wherein games failed to achieve all their intended effects (e.g. Clark, Tanner-smith, and Killingsworth, 2014; Morris et al., 2013; Wouters et al., 2013). At this point, it is still not clear how best to use games for educational purposes. It is difficult to generalise and compare studies to resolve the mixed results for studies using very diverse methods.

We currently lack a full scientific understanding of how and why some games succeed or fail to engage players to impart learning. For example, there have been perplexing findings of physiological indications of positive affect after failing a game (Ravaja et al., 2005; Ravaja et al., 2006; Salminen and Ravaja, 2008), suggesting players actually enjoyed failing. Such could defeat the entire purpose of a serious game if players enjoy losing instead of winning due to a flawed game design that fails to motivate players in the intended way. This phenomenon might be a critical problem in, for example, an educational farming simulation game in which the designers had accidentally made diseases more interesting to play with than building a successful farming business, resulting in players just spreading plague among their
livestock and destroying the farm instead of learning the intended lessons about agriculture market pricing choices. Similarly puzzling is that many players teach themselves complex games in their leisure time, and yet people often struggle with much less complex systems in the lab (Fischer and Gonzalez, 2015; Meadows and Wright, 2008; Sterman, 2000). Paradoxes such as these highlight the need for a common framework within which to study these vexing phenomena.

The primary problem being addressed by this research is: The mixed results of educational game research and the lack of a unifying framework that is able to specify the broader relevance to educational game design of narrower studies on specific phenomena.

1.2 Research Questions

This research aims to improve the design of educational games. Games have been developed for a variety of positive benefits and learning outcomes. However, progress has been slow, methods diverse, and results mixed. Designing educational games would be considerably easier if there were an established body of scientific principles and models describing how engagement and learning work in games. The general focus underlying this research is how to establish what specific properties of game systems have what specific effects on outcomes such as engagement and learning. This is refined into the following more specific research questions:

1. The Primary Research Question: How can serious games research move from proof-of-concept studies towards building up a formal model of how specific design changes affect specific outcomes such as learning and engagement?

2. How can different combinations of game features interact to cause players to engage in off-task behaviour instead of trying to win the game (and thereby likely miss the purpose of an educational or serious game)?
3. What is the relationship between the complexity of a game, and the player’s ability to master it? What form of complexity best arouses the curiosity of players?

This thesis focuses on addressing the primary research question (RQ1), with the two secondary research questions (RQ2 and RQ3) being supplementary examples with which to explore RQ1. Thus RQ2 and RQ3 are not the primary focus of this research, and instead serve as practical examples to apply and demonstrate the primary contribution of this research: Using the proposed framework to undertake empirical experiments demonstrates its value and possible ways it can be applied to educational games research processes. Therefore, this thesis reports on two empirical studies - one investigating RQ2, and one investigating RQ3.

This research proposes a framework to address the first, primary research question, and demonstrates its application in empirical experiments to address the two other, RQ2 and RQ3. The primary research question is directly addressed by the main contribution of this research, but RQ2 and RQ3 concern deeply complex phenomena that will take many empirical studies and theoretical developments to fully understand. They serve this thesis as practical examples to which the framework developed for RQ1 can be applied. This research will not completely solve, but will contribute toward an eventual solution to RQ2 and RQ3, particularly by laying the groundwork for future studies with the proposed framework to address the primary research question, RQ1.

1.3 Aims

This research has several main aims:

1. This research hopes to improve the design of educational games by providing a unifying framework that specifies the broader relevance of narrower studies of specific phenomena, thereby addressing RQ1.
INTRODUCTION

This framework will form the major contribution of this work and serve as the foundation for achieving the rest of the aims listed below.

2. Develop methods to objectively quantify several properties of games and their features that are relevant to outcomes of engagement and learning, suitable for use in the proposed framework and others like it.

Achieving aim two is also instrumental in achieving all the other aims. In order to determine if making game design change X results in outcome Y, one first needs to confirm that X was changed, and by how much. Therefore, this research also sought to develop ways to measure relevant properties of games, such as level and type of complexity (e.g. Kolmogorov complexity, Prokopenko, Boschetti, and Ryan, 2008), and challenge. Such measures would prove invaluable in achieving all other listed aims and answering the research questions, as well as proving to be useful tools to the broader field of games research.

These contributions (the framework addressing aim one, and objective measures of aim two) are then used to study specific phenomena to help establish connections between specific game features or properties, and specific outcomes of engagement or learning. This will provide some initial findings with practical applications for designing educational games. Specifically, these form aims three and four:

3. Identify problematic combinations of game features that can cause off-task behaviour, addressing RQ2.

Achieving aim three not only demonstrates how the framework can be applied to study a specific phenomenon, but also reveals complex interactions between game features to produce unintended consequences in terms of player behaviour, which is particularly pertinent to educational games: It can defeat the entire purpose of the game if players pursue an activity that is irrelevant or even counter-productive to the intended lesson of an educational game.

4. Identify a measurable form of game complexity that influences player curiosity, addressing RQ3.
Aim four serves as another example demonstration of the framework’s application, this time establishing if certain objective measures of game complexity are relevant to the curiosity of players. Managing complexity to arouse curiosity is critical to educational games that want to maximise engagement as a means to learning.

The primary aim is to lay a foundation for a more formal and precise approach to educational games research. This research aims to provide the tools necessary to study educational games in much finer detail, and to integrate the results of such studies to inform and contribute to the larger picture. The framework is designed to enable and encourage narrower, more detailed studies by precisely specifying their broader relevance of such work to related processes and outcomes. The framework aims to break down the process of how games achieve educational outcomes into more specific sub-processes that can be studied individually, and therefore simultaneously provide a structure that specifies how such narrower studies of individual sub-processes can be integrated to gradually build up a more formal, quantitative picture of the relationships between various design changes and various outcomes such as engagement and learning, which will ultimately inform educational game design.

This thesis proposes that educational game research can be divided into smaller subprocesses that can be studied in greater depth while still retaining relevance to the larger picture of how to design effective educational games, when that research occurs within a unifying framework.

1.4 Approach

By its nature, this research project is interdisciplinary, combining theories and methodologies from diverse fields such as cognitive science, complexity theory, and motivational psychology. Both the research questions and the aims determine what would be a suitable research approach. They require an approach that can disen-tangle the causal mechanisms at play and help to establish general principles of this phenomenon that can be useful in educational game design. Software engineering is
a rigorous and systematic approach that applies scientific and technological knowledge and methods to the design of software. This makes its methods particularly appropriate for this research topic.

In particular, software engineering often entails measuring the objectively quantifiable properties of software as a mathematical system, and the development of models to describe that system. Therefore, this research entails developing methods to objectively quantify properties of game systems, including complexity and challenge. Studying objectively quantifiable properties of games will assist with the rigour needed to tease apart causal mechanisms. This allows us to draw on a vast body of work in complexity theory to use measures of complexity to develop a model of how people learn complex systems. Using objective measures of game properties (such as complexity) will also help with the generalisability needed to be useful to the field of game design: if others measure similar properties for their games as those seen in this research, then we would predict similar results, but if they measure very different values, then a comparison may not be valid. Thus, it helps demarcate the area of applicability for this research.

As part of studying games as abstract mathematical systems, this research will use abstract games that do not represent or teach any particular subject matter. This improves internal validity by minimising confounding variables related to background knowledge and cultural associations with real-world topics. It can harm external validity by distancing this research from realistic educational game contexts, which generally are not abstract but actually do represent real-world subject matter. But this is combatted by the use of objective properties as explained above - If an educational game represents real-world subject matter (e.g. city planning), then as long as its measured properties are similar to those in this study, then a comparison should be valid.

Educational game research can be considered in terms of three entities and their relationships: the game, the player, and the real world (figure 1.1). These three entities are mediated by three processes: modelling the real-world domain into an appropriate game, the player mastering the game, and then the player transferring
that learning back to the real-world domain. These entities and processes form the core of the framework developed in this thesis, and formally defined and detailed in the chapter bearing the same name: The *Model-Master-Transfer* (MMT) framework for educational game research. MMT enables comparison and integration of diverse studies in the literature that might only focus on one of these entities or processes. In fact, one of the implications of MMT is that it will be much more efficient if research is able to study each of these processes and components individually to disentangle causal relationships and achieve greater depth of analysis. Such findings for individual, narrow studies can then be re-integrated into the larger picture to produce a more detailed and rigorously-tested model of how educational and serious games achieve positive outcomes.

Figure 1.1: Most educational game studies take a broad look at the entire process, from the game design, to the play and learning, to the transfer.

Most existing educational research bundles all these processes together into one process (figure 1.1), from the designing of a game based on a real-world activity, to the mastering of the game, to the transfer of that mastery back to the real-world domain (This common approach can be described as a proof-of-concept study). Consequently, if the results are negative, it is difficult to disentangle the cause of the failure, or at which step in the process the failure occurred. More precise answers can be gleaned by looking at these processes individually in greater detail. Therefore, as
enabled and recommended by MMT, this research will narrow its focus to one of the three processes to achieve greater depth and specificity.

As mentioned above, objectively quantifying relevant properties of games and their features is a necessary step to narrowing the focus to precisely study how specific game design changes affect outcomes such as engagement and learning. Developing such measurement methods is its own contribution to the field, and would fit within the first process of MMT, as part of the initial process of formally understanding the game in question before actually using it in the field or the lab (figure 1.2).

![Figure 1.2: The development of methods to objectively quantify relevant properties of games and their features would fall into the first process of MMT.](image)

To attempt to detail and experimentally examine all three processes (modelling a real-world domain as a game, the players mastering the game, and the players transferring that mastery to a real-world context) is too vast an undertaking to achieve here. Therefore this thesis has focussed on using MMT to study phenomena to do with playing and mastering the game. The process of mastering the game is absolutely critical to educational game research: If the player does not learn the game, then they cannot transfer that learning to the real-world. The process of mastery therefore deserves much in-depth, dedicated research. Thus, the empirical component of this research will narrow its focus to the process of mastering the game (figure 1.3) to achieve greater depth and precision in answering the research question and providing useful knowledge for educational game design. Specifically, the process of mastering the game will be investigated in two studies, addressing RQ2
and RQ3 described above, pertaining to the process of mastering the game: The first study will investigate how different combinations of game features can interact to cause off-task behaviour in players, and the second study investigate the relationship between game system complexity and engagement in terms of curiosity.

Figure 1.3: The focus of the empirical component of this research is the process of mastering the game, in order to achieve greater depth.

MMT, software engineering, and quantitative empirical research, all encourage developing models of the phenomenon under investigation. Developing a predictive and explanatory model facilitates the aim of establishing general laws and principles that can be useful in educational game design. It is complementary to many of the above methods. For example, a model can make use of objectively quantified variables once they are measured to generate predictions. MMT enables and encourages more detailed models of more specific phenomena to be integrated with each other to contribute to the larger picture of how educational games work. Therefore, the Dynamic Probability Response (DPR) model of challenge was also developed in this research (formal definition and full details provided in the chapter Study 1: Evaluating Play Appeal Factors & Engagement) to provide both an operational definition of the degree of challenge in different conditions, and provide predictions of engagement that will result from those different degrees of challenge.
The software engineering approach described above is quantitative. With the addition of empirical experiments, it is quantitative empirical research, which is an approach focusing on minimising subjective bias to establish general laws and principles of the phenomenon under investigation. The randomised controlled trial (RCT) is a prime example of this approach, used for its ability to isolate variables to disentangle causal mechanisms. An RCT includes both a treatment group (e.g. where learners are given the serious game to play), and a control group (e.g. where learners are given traditional classroom instruction, or a non-educational commercial game to play), and which individuals are assigned to either the treatment or the control group is randomised to minimise self-selection bias. Therefore, this research uses RCTs to test hypotheses to help establish general laws and principles that will be useful in educational game design. The approach taken in this thesis is outlined in figure 1.4.
Figure 1.4: The approach taken in this research, to address the primary research question through abstract modelling, then narrow focus to one process wherein RQ2 and RQ3 serve as example applications of the framework developed for RQ1, and using randomised controlled trials to explore RQ2 and RQ3. Boxes in bold are research contributions.
This will require creating a small game that allows the experimenter to manipulate relevant variables such as its degree of complexity, and records data during play that can be used to assess how well the game system was learned, and its overall ability to engage the player and provoke play. This quantitative empirical approach to the research is well suited to answering the secondary research question of how the specific game properties such as complexity relates to outcomes such as player behaviour, engagement, and mastery.

The interdisciplinary nature of this research is necessitated by the complexity of the research questions. Combining the theories and methods of various disciplines ensures a more well-rounded picture of the phenomena can be considered, and means that this research provides a unique perspective.

1.5 Contributions

This research has resulted in multiple contributions to this field:

1. Proposed Model-Master-Transfer as a formal framework to deconstruct serious games research into narrower sub-processes to be studied individually, and by which such narrow studies can be integrated into a larger mathematical structure.

Multiple meta analyses and literature reviews (Boyle et al., 2016; Clark, Tanner-smith, and Killingsworth, 2014; Connolly et al., 2012; McClarty et al., 2012) have found serious games research difficult to interpret, and have therefore recommended a shift away from proof-of-concept studies to narrower investigations of how specific design changes affect specific outcomes such as engagement and learning. Such studies could establish principles invaluable for serious game designers to improve the effectiveness of their games.

I propose a formal framework to facilitate that approach to serious games research - a framework by which to deconstruct serious games research into narrower sub-processes that can be studied individually in greater depth, and by which
such narrower studies can be integrated into a larger mathematical structure to contribute to the larger picture of how serious games achieve positive outcomes.

The MMT framework has many implications and applications not just for researchers, but also for practitioners such as game developers and educators. These contributions are detailed in a dedicated section. In summary:

A. Educational game developers can and should measure objectively quantifiable properties of their games that have relevance to theories of play and learning.

B. Educators using games should not just compare a game to traditional instruction, or to a completely different game, but compare the results of using the game with certain parameters adjusted up or down, using theories of play and learning to make predictions about outcomes of engagement and learning based on those game parameters.

C. Educators should not just measure the final outcome of performance on an academic test or transfer to a real-world task, but also measure and report how much mastery occurred within the game in order to determine if improvement in the game is proportional to improvement in the real world.

D. Educators can use MMT to identify early and terminate interventions that are unlikely to succeed in order to save time to try more promising interventions.

E. Researchers can use MMT to narrow the focus of their study to just one relationship between two or three variables (e.g. game system properties and outcomes of learning or engagement) to more precisely determine that relationship and test different theories of play, learning, and transfer individually. MMT can then be used to re-integrate the processes of play and transfer.

Another contribution takes the form of proposing methods to objectively quantify properties of games and their features.

2. MMT is elaborated in terms of how to model games in order to objectively quantify their properties, proposing Dynamic Causal Nets as a means to derive a measure of Kolmogorov complexity.
INTRODUCTION

These methods of modelling and calculating complexity can be directly applied in future studies, or serve as an example of how to achieve this approach to games research to help studies make use of MMT.

Next, the application of MMT to study specific phenomena is demonstrated. The two studies addressing RQ2 and RQ3 provide value in assessing hypotheses pertinent to educational game design, and as such will help develop our theories of play motivation interactions, the relationship between game complexity and curiosity, and therefore ultimately our understanding of engagement and learning in games.

3. Study 1: RQ2

A Developed a predictive model of challenge that provides methods to objectively quantify the degree of challenge present in different conditions in games, allowing these measurements to be used in MMT or similar frameworks.

B Found support that a challenging alternate activity can distract players from pursuing an activity visually marked as positive. This suggests designers should not rely on such communicative cues and need to ensure that the win condition is more challenging than other, disruptive activities.

C Found support that certain forms of feedback for pursuing an alternate activity can distract players from pursuing an activity that is both challenging, and visually marked as positive. This suggests designers may want to dampen feedback received from failure and/or increase feedback received from success.

D Did not find support for a simple, linear interaction between these game features in terms of engagement outcomes. This highlights the lack of work on theories of how play motivations may interact, distract, or amplify each other.

4. Study 2: RQ3
A. Supported the hypothesis that learning and engagement co-occur at a certain level of complexity - but only implicit learning (not explicit).

B. Found significant interactions between individual trait curiosity and explicit measures of engagement. It suggests educational game designers should cater for low trait curiosity individuals because high trait curiosity individuals were more interested regardless of complexity level.

C. Suggests that maximising engagement and implicit learning may require designing games with certain levels and kinds of complexity.

Additional contributions to extend and elaborate MMT are part of ongoing work included in the appendices:

5. An expansion of MMT to incorporate additional, social variables is part of ongoing work, included in appendix A. This work details the factors that affect how much a game is able to spread across a population to multiply its benefit to society.

6. A formal model of complex games as a cognitive habitat that gives rise to a diverse ecosystem of competing strategies is included in appendix B. This model provides a much finer level of detail in how mastery may develop in MMT.

This body of work highlights the utility of more detailed models of how specific properties of game systems can be used with player traits to predict outcomes such as engagement or learning, and that more work is needed on such models of these complex interactions between variables. The findings of the empirical experiments, combined with the MMT framework, pave the way for future research testing the interactions of the variables found to impact outcomes such as engagement and learning.

By adopting a rigorous quantitative empirical research approach that tests the relationship between objectively quantifiable properties of game systems (such as complexity) and outcomes such as engagement and learning, this research will contribute novel findings to the literature.
INTRODUCTION

This research is equally valuable to both educators and game designers. It provides general information on both how to teach, and how to keep people engaged, whether the context is an educational game, a traditional classroom, or a pure-entertainment commercial video game. Aside from its practical utility, this research also has academic value in contributing a framework to break down the process of learning through educational games, and in so doing specify the broader relevance of narrower research. This demonstrates the value of an interdisciplinary approach to complex research questions.

1.6 Structure

The structure of this thesis is shown in figure 1.5. The thesis is organised into the chapters according to the problems identified above. First is a literature review to provide most of the necessary background. Subsequent chapters detail the major contributions of this research in addressing the research questions explained above.

The first chapter following the literature review proposes the MMT framework to address the primary research question. The following chapter details how one could model games and subsequently calculate their Kolmogorov complexity, as would prove useful in MMT. This is followed by study one to explore RQ2 - how game features may interact to produce off-task behaviour. Study two is then presented to address RQ3 - how game system complexity relates to engagement through curiosity. Finally the discussion and conclusion to tie together the studies, what they mean for MMT and serious games research, limitations and future directions, and various practical applications for game designers and educators.

Additional contributions involve ongoing work to expand and elaborate MMT in more specific directions. These are included as appendices, with the first detailing factors of social context that affect how well a game may spread through a population to reach more individuals, and the second considering how game complexity and emergence could be defined as the level of diversity in an ecosystem of competing strategies.
Figure 1.5: An overview of the structure of this thesis, organised by the primary and secondary problems and questions detailed above. Major sections shown in boxes, with the darker, bold boxes showing the major contributions of this research. Reading from top to bottom of the diagram shows the order in which the reader will encounter each section.
Chapter 2

Literature Review

This research project was largely motivated by mixed findings in the literature on educational games and learning about complex (particularly emergent) systems. This required gathering together material from diverse fields: complexity theory and the philosophy of science for definitions and models of emergence, cognitive science and education literature on teaching about emergence (including both theoretical and empirical papers), and surveying the game design literature for discussion of emergent game designs and exploratory play. Many of these fields have produced developments relevant to each other, but communication across these fields is unfortunately uncommon. Bringing these diverse perspectives together can provide a new perspective on the primary and secondary research questions. Gathering together and comparing the parallel research in diverse fields provides a fuller picture of what is known about this phenomenon of learning complex games.

The following literature review uses different disciplines to explain how each element of this research topic (e.g. learning, play, complexity) can be conceptualised in relation to the primary and secondary research questions. In doing so, this literature review discusses how the different scientific disciplines complement each other, contribute to answering these questions, and ultimately provide some convergent predictions on how to design inherently learnable systems.
This research occurs at the intersection of multiple disciplines, and their unique perspectives on this topic are explained in this review. First is a broad overview of relevant research on learning about systems and educational games, which demonstrates the need for this research. This includes a summary of relevant theories of how learning occurs in games, and how choice of theory would impact the focus of game design and research. A review of multiple meta-analyses of educational and serious games demarcates the primary problem addressed in this research - the diversity of methodologies and results making the literature difficult to interpret, and need for research establishing generalisable principles of how specific game design changes cause specific outcomes such as engagement and learning. Existing frameworks are then collected and reviewed to extract the primary entities and processes involved in how educational games generate learning.

After this literature on the primary problem is a review of related literature on the two secondary problems. Literature in the psychology and design of games and play suggest some hypotheses regarding both RQ2 and RQ3, revealing possible ways that play motivations could come into conflict to derail a game’s goal, and ways that complexity could be used to provoke curiosity to create an inherently learnable game system. Finally, the literature on complexity theory is compared to complementary literature on human cognition to derive an hypothesised inverted U-shaped relationship between complexity and learning to address RQ3.

2.1 Learning

2.1.1 System Dynamics & Complex Problem Solving

Games are fundamentally mathematical systems (Adams and Dormans, 2012), and can therefore be analysed using tools from complex systems theory. Study of learning about complex systems often involves the use of microworlds: interactive software that simulate a real-world system or an arbitrary mathematical system, that allows researchers to precisely control the structure and dynamics of the system that contribute to complexity.
System dynamics is the modelling of phenomena and problems to identify and test solutions, and to study how well people are able to solve problems within such models (Sterman, 2000). At the core of such fields of systems research is the proposal that fundamentally, anything can be modelled as a system and that therefore similar principles and patterns can apply across systems with similar structures, even if they are composed of radically different things (such as comparing cardiovascular systems to city traffic systems using fluid dynamics). This is called isomorphism: When two systems consist of the same formal structure they are said to be isomorphic.

In parallel, very similar research has been conducted as a separate field of research: Complex problem solving (CPS). CPS is often defined as "[T]he successful interaction with task environments that are dynamic (i.e., change as a function of the user’s interventions and/or as a function of time) and in which some, if not all, of the environment’s regularities can only be revealed by successful exploration and integration of the information gained in that process." (Frensch, 1995). The majority of CPS research is on identifying the factors that contribute to complexity, and discovering the conditions and strategies that help overcome that complexity (for reviews see Frensch, 1995; Funke, 1995; Funke, 2012; Liu and Li, 2012; Quesada, Kintsch, and Gomez, 2005).

In both fields, one of the most striking and consistent findings is just how bad people are at solving problems within relatively simple dynamic systems (Meadows and Wright, 2008; Rouwette, Größler, and Vennix, 2004; Sterman, 2000). For example, people consistently underestimate the future magnitude of a variable that grows exponentially (Sterman, 2000). “Stock and flow failure” is the term for the pervasive inability to infer the behaviour of dynamic systems (Fischer and Gonzalez, 2015). This has highlighted the need for research into improving people’s understanding of complex systems.
LITERATURE REVIEW

2.1.2 Education via Games

Games have received a lot of attention recently as powerful self-motivating systems for transferring positive benefits. Games for learning purposes are merely a subset of the broader category of serious games. According to Blumberg et al. (2012) serious games are defined as “generally designed to entertain and educate players, and promote behavioral change via the incorporation of prosocial messages [...] embedded within game play” (p. 334). To that end, it is a broad category of games designed for any positive outcome beyond the game itself, such as facilitating or encouraging exercise, citizen science, or social activism (Adams and Dormans, 2012; Barton et al., 2016; Blumberg et al., 2012; Bogost, 2007). Gamification refers to approaching the problem from the opposite direction: Taking an activity and adding some game elements to it, such as points, trophies or leaderboards, to try to encourage certain behaviours (Barata et al., 2014; Deterding, 2012; Morris et al., 2013). The line between a gamified serious activity and a serious game is somewhat blurry, depending on how many game elements or serious elements are retained in either case.

The possible intended outcomes of serious games are diverse. Connolly et al. (2012) conducted a systematic literature review and categorised serious game purposes as being either affective and motivational, behaviour change, knowledge acquisition/content understanding, motor skills, perceptual and cognitive, physiological, social/soft skill outcomes, and various. Similar categories were used in an update to that review (Boyle et al., 2016). While potential applications for serious games are broad, these reviews found that the majority could be categorised as “games for learning”. Therefore, while educational games are a subset of serious games, they are very prominent, which makes them a highly relevant subset.

In this thesis I focus on educational games as a highly relevant domain that combines the use of technology with educational games. To that end, I will only focus on examples of that domain even though my proposed framework could potentially be used more broadly in relation to other serious games in non-electronic formats.
Educational video games are a relatively new phenomena that use technology to support educational games (Gee, 2005). Gee explains some of the ways the technology offers support, such as programming the game to offer relevant feedback and information just as it is needed. He uses the example of System Shock 2\textsuperscript{1}, which includes explanations within devices that are each found near to where the information will prove useful to make progress in the game.

Given games’ emphasis on systems and interaction, it is not surprising that they have received much attention teaching science subjects (e.g. Cheng and Su, 2012; Squire et al., 2004). Researchers have also examined the capacity for games to teach cognitive skills such as problem solving (Shute, Ventura, and Ke, 2015), spatial skills (Green, Pouget, and Bavelier, 2010; Shute, Ventura, and Ke, 2015), persistence (Shute, Ventura, and Ke, 2015; Ventura, Shute, and Zhao, 2013), metacognition (Kim, Park, and Baek, 2009; Moncarz, 2011), and cognitive adaptability (Gallagher and Prestwich, 2013). Whatever the game hopes to teach, there must be some theoretical mechanism by which it happens.

2.1.3 Implicit vs. Explicit Theories of Learning for Games

An often-cited thinker on this question is Gee (2005a; 2009), who proposes multiple mechanisms by which games might promote learning, such as helpful timing of information, or providing concrete first-hand experiences to make recall easier. Various learning theories can be applied in the design and evaluation of educational games and more specifically educational video games. The choice of theory puts emphasis on different aspects of the game’s design. Many learning theories can be considered in terms of whether they describe implicit or explicit learning - a distinction endorsed by dual process theories of cognition.

Dual process theories are a family of models of cognition that split knowledge and learning into two processes: type 1 and type 2, or implicit and explicit (e.g. Evans, 2012; Harteis et al., 2012; Kahneman and Klein, 2009; Osman, 2004; Smith

\textsuperscript{1}System Shock 2 is a first-person action role-playing survival horror video game for personal computers. It was published in 1999 by Electronic Arts. (https://www.imdb.com/title/tt0308956/)
These models attempt to explain, for example, how knowing how to ride a bike, and knowing the physics of riding a bike, are not one and the same. Or more broadly, how one could select a correct answer based on intuition (a “gut feeling”), or select a correct answer based on careful reasoning.

There is no consensus on what exactly implicit and explicit knowledge are, or if they are better conceptualised as two extremes of a single continuum (Osman, 2004). But there is commonality in how they are defined operationally: implicit knowledge is elicited by getting the person to perform or react, and explicit knowledge by getting the person to describe or diagram what they know or what they plan to do (Assaraf and Orion, 2005; Brandstädt, Harms, and Großschedl, 2012; Charles and Apollonia, 2004; Geddes and Stevenson, 1997; Greiff et al., 2015; Kim, 2015; Künsting, Wirth, and Paas, 2011; Lücken and Sommer, 2010; Osman, 2008; Pretz and Zimmerman, 2009). These would be ways of operationalising the learning gains in educational games. However, these methods are not without debate – Frensch and Rünger (2003) give an overview of such measures and suggest that dual-process theories need to focus on modelling what the two processes are, rather than just measuring them.

There are several possible models. Causal Bayes nets (CBNs) have emerged as highly effective models of how people develop explicit knowledge of how novel systems work (Gopnik, 2011; Gopnik and Schulz, 2007; Gopnik and Wellman, 2012; Schulz, Standing, and Bonawitz, 2008; Sobel, Tenenbaum, and Gopnik, 2004). This learning theory posits that learners discover causal relationships in systems (e.g. how relative air speed and density affect the drag forces on the rudder necessary for steering the aircraft) via intervention (e.g. exploratorily trying different flight manoeuvres at different altitudes and speeds), and thereby build a mental model. Such a mental model can simulate what would occur in novel situations never before encountered. As such, causal Bayes nets can be used to formally represent mental models.
A CBN is an arrangement of nodes with links representing causation. These CBNs are used to represent the actual real world system, and also to represent the knowledge of a potential learner (the “belief net” or mental model) – which would start relatively empty and grow to (hopefully) match the CBN of the actual system.

In a CBN, each node represents a variable (for example, a “weather” node could be rainy or sunny), and each link represents a causal influence (for example, the “grass” node could be dry or wet, depending partially on the weather). CBNs are very useful for trying to assess the probability of a certain state, given other states. For example, one might want to know, “what are the chances that the grass is wet, given the chance of rain today?”, or, “Given that the grass is wet, what are the chances that it rained today?” Specifically, there may be a 20% chance of rain, which changes the chances of the neighbour turning on their sprinklers from 40% down to 1%. Both rain and the sprinkler affect the chances of the grass being wet (figure 2.1). Causal Bayes nets are named so after Bayes Rule ($Pr(B|A) = Pr(A|B) \times Pr(B)/Pr(A)$), which explains how to calculate the probability of an event given another event. There is no shortage of resources explaining the details of CBNs (see e.g. Gopnik and Schulz, 2007), and so this broad overview will suffice here to discuss how it could influence educational game design and research.

![Figure 2.1: A simple example Causal Bayes Net](Bayesian Network 2020).
Designing or evaluating an educational game based on a theory such as CBNs puts the emphasis on abstract causal structure (Gopnik and Wellman, 2012; Ross, 2013; Steyvers et al., 2003): Do the causal relationships players have mapped in their mental models match up with the actual causal relationships in the game, and do the causal relationships in the game correspond accurately to the causal relationships that exist in the real-world domain? This is a very different design/research focus one would have than if one were to use an implicit learning theory, such as instance-based learning theory.

*Instance-based learning theory* (IBLT) posits that skills develop as learners associate a particular perceptual cue (e.g. falling downward) with a particular motor response (e.g. pulling up on the joystick), building up an instance library from which an action can be automatically retrieved when faced with a similar situation as experienced in the past (Fischer, Greiff, and Funke, 2012; Geddes and Stevenson, 1997; Gonzalez and Quesada, 2003; Logan, 1988). This is contrasted with explicit rule-based reasoning, where one thinks through the logical consequences of possible choices to determine which one is best (Jong, Hoog, and Vries, 1993; O’Hara and Payne, 1998; O’Hara and Payne, 1999; Osman, 2010; Sedig and Haworth, 2014). It therefore has relevance for skills that can be performed in real-time, such as piloting an aircraft. Educational video games based on IBLT are likely to shift the focus from abstract causal structure to the perceptual fidelity of the stimulus the game provides (e.g. the graphics and sound), and the physical fidelity of the control system (e.g. the joystick), to ensure players build an accurate instance library through game experience.

This research shall remain agnostic as to the actual mechanism behind implicit or explicit reasoning where possible. Whatever theories may ultimately prove reliable, it would clearly be useful for a unifying framework for educational games to be able to accommodate studies that want to use vastly different theories of learning, or combine some measures of implicit and some measures of explicit learning. These learning theories (CBNs and IBLT) will serve as useful examples to refer to throughout this research.
2.1.4 Ambiguity of Educational Game Outcomes

The primary problem addressed by this thesis is the mixed results and lack of unifying framework in this area: Educational games research has revealed ambiguous outcomes. While some studies were able to use educational video games to achieve measurable benefits (Ibrahim et al., 2012; Morris et al., 2013; Squire, 2008), others have had mixed results including weak effects or finding support for only a few of their hypotheses (Clark, Tanner-smith, and Killingsworth, 2014; McClarty et al., 2012; Wouters et al., 2013). Others still have been unable to achieve their intended outcomes altogether (Frank, 2011; Ronen and Eliahu, 2000). Reviews and meta-analyses paint a similar picture. For example, Powers and Brooks (2014) conducted a meta-analysis of video game training transfer and found a mixture of mostly small to medium effects, depending on game type and outcomes measured. Sala et al. (2018) found negligible to no benefits for motor skills from educational video game training in their meta-analysis. In contrast, Wouters et al. (2013) found in their meta-analysis that educational video games were more effective than traditional instruction for learning and retention (but not motivation), and Parisod et al. (2014) found that educational video games could be effective for several children’s health applications.

McClarty et al. (2012) reviewed the literature of video gaming in education and concluded that, "results are mixed because of differences in definitions and methodologies,” continuing to say that a more coherent research approach should investigate the content, structure and mechanics of educational games, creating definitions and models of the relevant properties of games.

Clark et al. (2014) tested multiple hypotheses in their meta-analysis and found support for some, such as that scaffolding is helpful and that schematic games rather than cartoon or realistic games are more effective. They conclude that research should shift from proof-of-concept studies that focus on medium, to studies that test how theoretically-driven design decisions affect learning outcomes.

Connolly et al. (2012) and Boyle et al. (2016) gathered various serious games studies and summarised notable positive and null findings. They found the studies were very diverse with respect to their underlying theoretical models,
methodological approaches, and outcomes reported, noting a majority were qualitative study designs. Despite learning being the most common purpose of the studies reviewed, they found that, “Evidence that games lead to more effective learning was not strong” (Connolly et al., 2012, p. 671). They conclude that, “more RCTs [randomised controlled trials] should be carried out” (Connolly et al., 2012, p. 671), and, “future research will benefit from detailed experimental studies that systematically explore which game features are most effective in promoting engagement and supporting learning” (Boyle et al., 2016, p. 188).

The above meta-analyses and literature reviews recommend research shift from proof-of-concept studies to more systematically investigate how specific design changes to the properties of educational video games can affect outcomes such as learning and engagement.

Throughout this literature review, most proof-of-concept studies cited exemplify a logical sequence of steps that is generally left implicit. Discussing the relevance of the literature can be clarified (in terms of relating game properties to outcomes of learning and engagement) by making that sequence explicit, and so I shall. For an educational game to achieve learning hinges on the success of a multi-step process exemplified (but implicit) in many of the proof-of-concept studies cited:

0. The student starts with insufficient knowledge/skill;
1. plays the game;
2. develops mastery of the game;
3. and is then able to transfer that mastery to the context of a real-world task/problem.

Stage two is crucial to this entire endeavour, and yet it has received very little attention within the educational games literature to how exactly players learn to master a game. If one cannot achieve that step in the process, then one has no reason to believe one can improve their performance on the real-world task (one cannot have transfer of skill that one has not even acquired yet). Therefore, stud-
ies dedicated to such a narrow, singular step of the process are of vital importance, but they also need to be able to be integrated with the larger process of educational game usage (e.g. relating mastery to transfer). It could remove ambiguity when a research design is narrowed down to more deeply investigate a singular relationship between, for example, engagement and the degree of agency afforded in a game. However, such a study could be easily overlooked by educational video games researchers as irrelevant for lacking a test of any learning at all.

According to the meta-analyses reviewed above, narrow studies on causal relationships between specific educational video game properties and resulting experience or behaviour should be able to contribute to the larger picture of how to design and apply educational video games to achieve beneficial outcomes. This could be encouraged by a framework that specifies their relevance by providing a means to integrate studies of narrow sub-processes into the larger process of educational video games usage. Therefore, my framework’s goal is not only to decompose educational game research into more specific sub-processes for detailed study, but also to specify the broader relevance of narrower research in order to assist researchers and practitioners in piecing together disparate, narrow research to form a more comprehensive causal model of how educational and serious games can achieve positive outcomes for players. For example, a literature review or meta-analysis should be able to link together multiple narrow studies to form a causal chain that spans the beginning of the process (designing the game), all the way to the end of the process (transferring learning to the real world).

The recommendations of the above meta-analyses can be elaborated with additional suggestions from related research. For example: To ensure one’s findings are not just an artefact of using a specific educational subject, or the result of cultural associations and background knowledge, the focus should shift from the surface features to the abstract properties of games as systems. Semantic embedding (when a simulation depicts a recognisable real-world system) can harm knowledge acquisition by fostering a set of presumptions that may not be accurate and tend to go untested (Beckmann and Goode, 2014). Goode (2011) advises it is therefore best to create simulations that are abstract, so that they do not have a recognisable, familiar
real-world counterpart. By eliminating obfuscating variables related to the narrative and world of the game, and focussing on analysing quantifiable variables of the game as a dynamic mathematical system, one can ensure that the results of a study pertain to game systems generally, and are not just due to cultural associations and background knowledge of a particular domain. This sidesteps the common difficulty in trying to interpret game studies on teaching one topic (e.g. piloting aircraft) to try to figure out if they necessarily apply to games teaching different subject matter (e.g. agriculture). The approach of focusing on abstract game systems is the opposite of most educational game studies, which attempt to teach a specific subject or skill, and therefore the present research will contribute novel findings to address this gap in the literature.

This approach is analogous to the psychological study of music – music is ultimately a waveform, which has properties open to mathematical analysis. Where most other studies compare a game to a lecture, or compare two very different commercial games to each other, this research will compare a game to itself with only one variable about its underlying system changed. Similar methodologies have been used in software engineering to, for example, calculate the network complexity of software to help assess the software’s robustness (Wen, Kirk, and Dromey, 2007; Wen, Dromey, and Kirk, 2009). The major difference here is that instead of robustness, the outcomes of interest are engagement and learning.

This research project aims to bridge the gap between the practical case studies of specific educational games, and the high-level theories of cognition (figure 2.2). Therefore, to maximise the value and generalisability of the research, research needs to:
1. narrow the focus to how a specific design change affects a specific outcome such as engagement or learning. One place to start is to dig into this neglected stage two process – how players master games,
2. focus on quantifiable properties of the game systems,
3. and keep the game systems independent of specific domain knowledge (i.e. avoid semantic embedding).

This review demonstrates the primary problem addressed by this thesis. According to the meta-analyses reviewed above, narrow studies on causal relationships between specific game properties and resulting experience or behaviour should be able to contribute to the larger picture of how to design and apply educational games to achieve beneficial outcomes. This could be encouraged by a framework that specifies their relevance by providing a means to integrate studies of narrow sub-processes into the larger process of educational games usage. Therefore, my framework’s goal is to specify the broader relevance of narrower research in order to assist researchers and practitioners in piecing together disparate, narrow research to form a more comprehensive causal model of how educational games can achieve positive outcomes for players. Many frameworks have previously been proposed, but tend to address very different problems to the research problem identified here.

2.1.5 Entities & Processes in Existing Frameworks

As mentioned above, multiple literature reviews and meta-analyses have suggested a shift from proof-of-concept studies to narrower investigations of specific design changes and their outcomes (Boyle et al., 2016; Clark, Tanner-smith, and Killingsworth, 2014; Connolly et al., 2012; McClarty et al., 2012). Therefore, existing frameworks for educational games were collected to extract the major entities and processes they consider pivotal in educational games. The purpose of this review process was to address the primary research question: How can research on game-based learning be broken down into more specific sub-processes that can be studied individually (to achieve greater depth and precision)? Having broken it into narrower topics,
how can narrow studies on more specific sub-processes then be used to build up a broader causal model of how serious games achieve positive outcomes? The aim was to identify the major entities and processes that are commonly assigned a causal role in models of educational and serious games.

Online databases (specifically Emerald, Eric, ScienceDirect, Scopus, Web of Science, and Google Scholar) were searched in the title and keywords fields where possible to minimise false positives (many serious games studies that do not propose frameworks nonetheless use the term at some point in the main body of the text) using the search term:

```
(("computer gam*" OR "video gam*" OR "serious gam*" OR "simulation gam*" OR "games-based learning" OR "educational gam*") AND ("framework" OR "model") AND ("caus*" OR "mechanism" OR "predic*" OR "variable"))
```

The focus was on models that posit causal relationships or mechanisms because this thesis’ aim was not to satisfy stakeholders or guide designers directly, but to assist at the more foundational level by improving our collective scientific understanding of educational and serious games. The collected papers’ abstracts were screened to trim away papers not relevant to this investigation: Those that did not propose a causal model of how games achieve positive outcomes for players. For example, those that merely modelled player enjoyment, or clients’ and stakeholders’ concerns about managing a serious game project, were excluded. To be included, studies had to:

1. propose and explain a framework,

2. pertain to learning in games generally (whether using a serious game, or a commercial entertainment game), not specialised only for a specific game genre (e.g. puzzle games) or only for teaching a specific topic or skill (e.g. piloting aircraft),

3. propose causal mechanisms, not simply organising categories,
4. be a general framework of learning in games, not a technique or notation for converting a game into a formal model.

These criteria trimmed the pool of papers from an initial size of 422 down to 25. When supplemented with relevant papers already gathered in background reading before this process, the total came to 29 papers proposing a total of 27 frameworks. These are summarised in table 2.1.

### Reviewed Frameworks Summary

<table>
<thead>
<tr>
<th>Citation</th>
<th>Framework Name</th>
<th>Summary</th>
<th>Use</th>
<th>Entities</th>
<th>Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbasi, Ting, and Jamek, 2015</td>
<td>Unnamed</td>
<td>Relates engagement to factors of attention, retention, reproduction, motivation, and learning.</td>
<td>Derive hypotheses based on causal relationships</td>
<td>Game, Player</td>
<td>Learning through play</td>
</tr>
<tr>
<td>Ak, 2012</td>
<td>Unnamed</td>
<td>Framework for evaluating games, including input factors such as the curriculum, an intermediate play/learning cycle, and finally outcomes such as motivation.</td>
<td>Guide evaluation of educational games</td>
<td>Real-world domain, Game, Player</td>
<td>Game design, Learning through play</td>
</tr>
<tr>
<td>Amory, 2007</td>
<td>Game Object Model version II (GOM II)</td>
<td>Organisational framework relating the game, player, the problem to be solved, and the social environment.</td>
<td>Guide design or evaluation of educational games</td>
<td>Game, Player</td>
<td>Learning through play, Social collaboration</td>
</tr>
<tr>
<td>All, Nuñez Castellar, and Van Looy, 2015</td>
<td>Unnamed</td>
<td>Used interviews with stakeholders and game designers to derive an evaluation framework including learning, motivational, and efficiency outcomes.</td>
<td>Guide evaluation of educational games</td>
<td>Game</td>
<td>Learning through play</td>
</tr>
<tr>
<td>Ariffin, Oxley, and Sulaiman, 2014</td>
<td>Unnamed</td>
<td>Causal relationship between factors of learner background and motivation, and motivation to performance.</td>
<td>Derive hypotheses based on causal relationships</td>
<td>Player</td>
<td>Learning through play</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Framework</td>
<td>Description</td>
<td>Guide design or evaluation of games</td>
<td>Planning focus</td>
<td>Application focus</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>Arnab et al., 2015</td>
<td>Learning Mechanics-Game Mechanics (LM-GM)</td>
<td>Relates specific learning mechanics to game mechanics.</td>
<td></td>
<td>Game, Learning through play</td>
<td></td>
</tr>
<tr>
<td>Carvalho et al., 2015</td>
<td>Activity Theory-based Model of Serious Games (ATMSG)</td>
<td>Relates the game, player, and social community to outcomes based on goals, tools and actions.</td>
<td>Guide design and evaluation of educational games</td>
<td>Game, Learning through play</td>
<td></td>
</tr>
<tr>
<td>De Freitas and Oliver, 2006</td>
<td>Unnamed</td>
<td>Framework to guide evaluation and design of educational games based on the context, learner characteristics, the pedagogy to teach, and the tool of media representation used (the game).</td>
<td>Guide design and evaluation of educational games</td>
<td>Game, Learning through play</td>
<td></td>
</tr>
<tr>
<td>Duke, 1980</td>
<td>Unnamed</td>
<td>An iterative stepwise process for designing a serious game for a client</td>
<td>Guide game design</td>
<td>Game, Game design</td>
<td></td>
</tr>
<tr>
<td>Eseryel et al., 2013</td>
<td>Unnamed</td>
<td>Relates motivations of interest, competence, autonomy, relatedness, and self-efficacy to engagement and problem representation</td>
<td>Derive hypotheses based on causal relationships</td>
<td>Player, Learning through play</td>
<td></td>
</tr>
<tr>
<td>Fjellingsdal and Klockner, 2017</td>
<td>Environmental Educational Game Enjoyment Model (ENED-GEM)</td>
<td>Relating factors of learner background and game properties to learning outcomes, via three-stage process starting with motivation, then gameplay, then learning outcomes.</td>
<td>Derive hypotheses based on causal relationships</td>
<td>Game, Player, Learning through play</td>
<td>Applying to the real world</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Year</td>
<td>Model/Concept</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------</td>
<td>---------------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gunter, Kenny, and Vick</td>
<td>2006</td>
<td>RETAIN</td>
<td>Integrates elements of the ARCS motivational model, Events of Instruction, and Bloom’s Taxonomy, into Salen &amp; Zimmerman’s (2003) Multivalent Model of Interactivity to derive guidelines for educational game design.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hainey, Connolly, and Boyle</td>
<td>2010</td>
<td>Unnamed</td>
<td>Evaluates game effectiveness based on the preferences, attitudes, perceptions, and motivation of the learner and instructor, and the environment in which it is used.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harteveld</td>
<td>2010</td>
<td>Triadic Game Evaluation</td>
<td>Macro cycle consists of three micro cycles: Engagement during play, making meaning or taking value from the game, and applying it to the real world.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jackson and McNamara</td>
<td>2017</td>
<td>Motivation and Mastery Cycle Framework</td>
<td>A cyclical causal framework where player factors and game factors affect initial interest, interest affects persistence, persistence affects mastery, mastery affects self-efficacy, and self-efficacy affects interest, creating a positive feedback loop.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jorge and Sutton</td>
<td>2017</td>
<td>FUNIFICATION Model 2.0</td>
<td>A Likert-type tool for evaluating the design of educational games.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kiili, Kiili et al.</td>
<td>2005; 2014</td>
<td>Unnamed</td>
<td>Relates factors leading to flow, signs flow has occurred, and outcomes such as schema construction.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author(s)</td>
<td>Year</td>
<td>Title</td>
<td>Summary</td>
<td>Guide Research</td>
<td>Player, Social Context</td>
</tr>
<tr>
<td>------------------------</td>
<td>------</td>
<td>-----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>----------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Klabbers, Unnamed</td>
<td>2018</td>
<td>Micro-cycle describes learning through play. Macro-cycle describes learning with external assistance such as debriefing.</td>
<td>Guide research Player, Social context</td>
<td>Learning through play</td>
<td></td>
</tr>
<tr>
<td>Landers and Armstrong, Technology-Enhanced Training Effectiveness Model (TETEM)</td>
<td>2017</td>
<td>Relates attitudes of students and teachers about technology to learning outcomes</td>
<td>Derive hypotheses based on causal relationships</td>
<td>Player Learning</td>
<td></td>
</tr>
<tr>
<td>Mayer, Unnamed</td>
<td>2012; Mayer et al., 2014</td>
<td>Framework for game evaluation to satisfy stakeholders, incorporating the various roles that can be taken (e.g. organisational, player, learner, professional).</td>
<td>Guide research Game, Social context Applying to the real world</td>
<td>Learning through play, Applying to the real world</td>
<td></td>
</tr>
<tr>
<td>Obikwelu and Read, 2012</td>
<td>Unnamed</td>
<td>Describes how constructivism can occur in educational games.</td>
<td>Guide evaluation of educational games</td>
<td>Player Learning through play</td>
<td></td>
</tr>
<tr>
<td>Plass, Unnamed</td>
<td>Homer, and Kinzer, 2015</td>
<td>Cycle of play driven by game features (incentive system, game mechanics, aesthetic design, narrative design, and musical score), looping through states of challenge, response, feedback, and back to challenge.</td>
<td>Guide game design</td>
<td>Game Learning through play</td>
<td></td>
</tr>
<tr>
<td>Roungas and Dalpiaz, 2016</td>
<td>Unnamed</td>
<td>Class diagram for an application for managing a design document for a serious game</td>
<td>Guide game design</td>
<td>Game Learning through play</td>
<td></td>
</tr>
<tr>
<td>Shi and Shih, 2015</td>
<td>Cognitive Behavioural Game Design (CBGD)</td>
<td>Relates game factors to outcomes of motivation</td>
<td>Guide game design</td>
<td>Game Play</td>
<td></td>
</tr>
<tr>
<td>Starks, 2014</td>
<td>Cognitive Behavioural Game Design (CBGD)</td>
<td>Relates certain goals of the game designer to certain game features, which in turn are associated with certain player outcomes</td>
<td>Guide game design</td>
<td>Game Learning through play</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.1: Summary of reviewed frameworks.

The frameworks summarised in table 2.1 share much commonality in terms of which entities and processes they included. Of the 27 frameworks collected, 19 (70.7%) included the game itself as an entity in the framework, and an equal number (19) also included the player in the framework. Seven (25.92%) included the real-world domain of knowledge to be taught, and only five (18.51%) included some kind of social context as an entity. Social context being the least common inclusion is perhaps due to the fact that, while all educational games involve games and players, not all require or even support multiple players. Therefore, including the game as an entity will be relevant to all educational games, but the social context will only be relevant to a subset.

In terms of processes, the vast majority of frameworks (24 or 88.89%) included the process of mastering the game. Four (14.81%) included the process of designing the game as part of the framework, and three (11.11%) included the process of transferring the mastery of the game back to the real world.

Many frameworks have been proposed for various processes and sub-processes of educational games, but few have similar aims to ours. Of the 27 gathered in above table, 12 (44.44%) serve as a game design guideline framework for satisfying clients and stakeholders when developing educational games at the start of the process.
Eleven (40.74%) aimed at evaluating the effectiveness of the game at the end of the process, seven (25.92%) were for deriving hypotheses based on causal relationships, and only two (7.4%) hoped to guide research on educational games.

However, none of these frameworks sought to address issues identified by the above meta-analyses and reviews (Boyle et al., 2016; Clark, Tanner-smith, and Killingsworth, 2014; Connolly et al., 2012; McClarty et al., 2012) such as mixed results, incomparability between studies with conflicting results, or facilitating research that moves away from proof-of-concept studies to investigations of how specific design changes affect specific outcomes. More specifically, none sought to provide a structure by which to break down educational games into smaller processes to study, or a structure in which to assemble narrow studies on specific causal relationships to build up the larger picture of educational games. In the next section I propose a framework to break down and specify such processes to get a clearer and deeper understanding of them individually, and how to then re-integrate those processes to form the larger picture again.

This review helped identify the major entities and processes that are commonly assigned a causal role in how educational and serious games achieve positive outcomes, and thus helped develop the framework proposed by this research to address the primary research problem. The meta-analyses reviewed above found educational game research needs to shift from proof-of-concept studies to narrower studies that investigate how specific design changes affect specific outcomes, highlighting the need for a model that can break down educational game use into narrower sub-processes to be studied individually, and also specify the relevance of such narrower studies to allow them to be integrated to gradually build up a more comprehensive model of how educational games achieve positive outcomes. Using this literature review of existing frameworks, a new framework is proposed in the below section, Model-Master-Transfer. This framework forms the first and primary contribution of this research project: A framework for deconstructing serious games research into narrower sub-processes to be studied individually, and for integrating such narrower studies into a larger mathematical structure.
But first there is another long-standing issue that needs to be discussed: How games and play relate to learning.

2.2 Game Design

2.2.1 Games vs. Simulations

A central issue in educational game research has been trying to integrate the learning aims with the actual game play. Failure to do so is an error akin to “dipping broccoli in chocolate” (Miller, 2013; Hall, Wyeth, and Johnson, 2014) - chocolate will improve the taste of broccoli no more than disconnected game elements will make a lesson fun. Bogost (2007) and Krajewski (2014) explain that play should not be a reward for enduring the learning of unrelated knowledge and skills, but that the knowledge and skills should be empowering tools, instrumental to success in playing the game. Games with their educational material built into the mechanics have been called conceptually integrated (Clark, Sengupta, Brady, Martinez-Garza, & Killingsworth, 2015). That is the enigmatic potential of educational games: just by having fun, people get the learning for free. Therefore, to make proper use of games to their full potential, this research must take into consideration the conditions where the learning is integral to the play.

To avoid the “chocolate on broccoli” problem, one must consider why games are more like chocolate than broccoli in the first place. What are games and why might they be useful in learning about complex systems? Many debates about the definition of a “game” have accumulated in academia (Costikyan, 2002; Crawford, 1984; Juul, 2005; Salen and Zimmerman, 2003). Popular definitions include, “a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome” (Salen and Zimmerman, 2003, p. 80), and particularly Juul’s definition:

A rule-based system with a variable and quantifiable outcome, where different outcomes are assigned different values, the player exerts effort
in order to influence the outcome, the player feels emotionally attached to the outcome, and the consequences of the activity are negotiable. (Juul, 2005, p. 84)

These definitions reveal that games are fundamentally mathematical systems, and can therefore be analysed using tools from complex systems theory. Adopting this viewpoint allows this research to draw upon a vast array of previous work on different kinds of systems using similar mathematical tools. Such tools enable the use of my framework to test hypotheses about the relationships between objectively quantifiable properties of games. Not just games, but virtually any real-world domain can also be modelled as a formal system, according to fields of study such as system dynamics.

Games and simulations cannot be infinitely complex, so how complex should they be? This is a tricky, long-standing problem in simulation theory, and is part of RQ3: asking if there is an ideal level of complexity for these educational games, where more detail would make it indecipherable, or less detail would make it boringly simple. There is no doubt that a serious simulation can focus too much on comprehensive realism and consequently suffer from being indecipherable to a “player” put before it.

Sauvé et al. (2007) sought specifically to distinguish between games and simulations. Which prompts the question: why do educational game researchers investigate games instead of simulations? Cook (2014) questioned the value of crafting all these definitions when all that matters to game designers is how to make games fun. This gets to the essence of games, and thus the core of scientific interest in games: They are worth investigating precisely because they are somehow fun. People do not play training modules. People play games. Cook (2014) calls games “endogenous systems of value”, emphasising their enigmatic ability to provoke people to care about what happens in them, and thus engage in play. That is the simple definition that will be adopted for this research: A game is a system that somehow incites play. How exactly they do so is not well understood, making designing games difficult.
2.2.2 Engagement

Play is a broad category covering diverse behaviours, and there are comparably multitudinous theories to account for all the different forms of play (for reviews, see Bateman, 2014; Hamari and Tuunanen, 2014). In many cases, proof-of-concept studies do not specify which theory or theories of play they are adopting for the design of their game (e.g. Barton et al., 2016; Miller, Dawood, and Kassem, 2012; Mohsen, 2016; Squire et al., 2004). It would be tremendously helpful for inter-study comparison if researchers could specify, for example, that they expected agency, challenge, or social bonding to be the primary source of engagement in their particular game, and designed with the intent of maximising that element of the game.

Once a researcher has selected one or more theories of play, they are able to investigate what efforts have been made to formalise the relevant variables: how to quantify the play appeal of the game. For example, agency can be quantified as the number of options available to the player at any given time, or as the probability of the player’s actions succeeding (Tornqvist and Tichon, 2019). There have also been various works proposing methods for quantifying the degree of challenge in a game (Aponte, Levieux, and Natkin, 2011; Fraser, Katchabaw, and Mercer, 2014; McMillan, 2013; Tornqvist and Tichon, 2019).

Objectively-measured game parameters have been neglected in the literature but are crucial to the development of gaming theory. Serious game design could be much more efficiently and effectively designed if there were precise models available based on objectively measurable properties of games.

There are multiple relevant constructs when it comes to measuring play: engagement, motivation, and affect. At its most basic, affect refers simply to the positive or negative valence of the experience (e.g. “are you enjoying yourself?”). Motivation is at a deeper level and could drive, for example, approach or avoidance behaviour (Elliot and Church, 1997; Elliot and Harackiewicz, 1996). Motivation and affect are related, but they are not the same thing as one can be motivated to complete something that they do not enjoy doing.
Engagement has been used to refer to the arousal or attention component of mood (Watson and Vaidya, 2003). However, here, engagement will be used in the literal sense of engaging with an item: attending to and physically or cognitively manipulating it. Games have the potential to be both physically and cognitively engaging. Note that this is a behavioural definition of engagement concerned with what one actually does, and therefore one can be engaged without necessarily being highly motivated or experiencing positive affect.

Affect and motivation are often causal factors of engagement as one is more likely to engage with an item if it is motivating, and more likely to repeat or continue engagement if it was enjoyable. But these are not required for engagement. For example, a student may be extrinsically motivated to engage with the lesson and may be extremely irritated, but engage nonetheless. Engagement is necessary for learning, regardless of how it is motivated or the accompanying affect. Of course, one can engage poorly with a lesson and consequently fail to fully understand it. Engagement can therefore have both a quantity and a quality, or a duration and a magnitude.

There are many theories of play, and determining which one is most applicable to a certain genre or type of game can have implications for the predicted engagement-over-time curve. Social play for the purposes of bonding and developing relationships is arguably more likely to result in an enduring form of engagement, as long as there are people to socialise with in the game. Contrast this with exploratory play relying on novelty to provoke curiosity. This would potentially result in a much shorter play duration due to a sharp drop-off in engagement once curiosity is satisfied.

The advantage of games for learning is that they (in theory) should not require an authority figure mandating engagement, but should be inherently engaging. To measure different motivations would be to investigate how games engage. To measure engagement would simply investigate if they do so, and how strongly. Different studies of games have measured engagement in different ways, some attempting to measure it indirectly via motivation or affect. It may therefore be ad-
visable to triangulate engagement, ideally using a measure from each category: a
direct measure of engagement, a measure of motivation, and a measure of affect.
This would further empower meta-analyses to decompose studies into these factors
of engagement, find trends and test more specific hypotheses.

2.2.3 Measuring Engagement

Affect is a dimension of both emotion and mood, the distinction being primarily
that emotion is assumed to consistently correlate with neurological states and phys-
iological responses and can be fleetingly temporary, whereas mood can last for days
and is not necessarily assumed to correlate with specific brain states or physiolog-
ical responses (Watson and Vaidya, 2003). Emotion is often measured in terms of
self-report, physiological responses (e.g. galvanic skin response, GSR, or electroen-
cephalography, EEG), or behavioural expression (e.g. vocal tone, facial expression,
body posture, or approach or avoidance behaviour). Mood usually relies just on
self-report tools such as the Current Mood Questionnaire (CMQ, Yik, Russell, and
Barrett, 1999), Positive and Negative Affect Schedule (PANAS, Watson, Clark, and
tellegen, 1988), Mood and Affect Checklist (MAACL, Zuckerman and Lubin, 1965),
Profile of Mood States (POMS, McNair, Lorr, and Droppleman, 1971), and Differen-
tial Emotions Scale (DES, Izard et al., 1993).

Physiological measures have been used to evaluate games (Nacke and Lind-
ley, 2010; Ravaja et al., 2005; Ravaja et al., 2006; Ravaja et al., 2008; Salminen and
Ravaja, 2008). Such measures have good temporal resolution, but attaching devices
to players’ heads, faces and fingers could interfere with their play experience. Mauss
and Robinson (2009) provide a detailed discussion of the accuracy and drawbacks
of different measures of emotion.

Self-report measures have proven very useful. The CMQ has proven reli-
able in measuring some dimensions of mood but not others (Watson and Vaidya,
2003). The MAACL and its revised version (MAACL-R, Zuckerman and Lubin,
1985) have also suffered from some problems of reliability and too much correla-
tion between what are supposed to be distinct negative mood scales - problems par-
tially attributed to its reliance on a checklist format, which is recommended against (Watson and Vaidya, 2003). POMS also suffers from poor discriminant validity of its negative scales, and the DES has the advantage of brevity, but at the cost of low reliability (Watson and Vaidya, 2003). PANAS is a frequently-employed tool that has received extensive study of its validity. PANAS was expanded into PANAS-X (Watson and Clark, 1994) to capture the more specific moods of all these other tools, however none of them share much agreement about what exactly each of these specific moods should be. In other words, they all conceptualise mood in terms of mostly different specific states. Even if a mood shares the same name in two systems (e.g. “anxiety”), they do not seem to be measuring the same thing. However, there is much more agreement in terms of the broader dimensions of mood: pleasantness (happy vs. sad), positive affect (excited vs. sluggish), engagement (aroused vs. still), and negative affect (distressed vs. relaxed). This research project’s aims are more focused on such broad dimensions, for which tools like PANAS are relatively reliable.

For measuring motivation, the intrinsic motivation inventory (IMI, McAuley, Duncan, and Tammen, 1989) is a ubiquitous measure, but there are questions around its theoretical foundations (Markland and Hardy, 1997). Continuing the example of a serious game designed to achieve education, one could focus more specifically on the motivation of curiosity. There are many tools for curiosity measurement (for a review of these measures, see Grossnickle, 2016), such as the Curiosity and Exploration Inventory Mark II (CEI-II, Kashdan et al., 2009), State–Trait Personality Inventory (STPI, Spielberger et al., 1979), Melbourne Curiosity Inventory (MCI, Naylor, 1981), Sensation Seeking Scale (SSS, Zuckerman, 1979), Novelty Experiencing Scale (NES, Pearson, 1970), Academic Curiosity Scale (ACS, Vidler and Rawan, 1974), and State Epistemic Curiosity Scale (SECS, Leherissey, 1971). The majority of these are mostly or solely for measuring curiosity as a stable character trait, not as an experienced state. The reliability and validity of the trait components of these tools have been much more rigorously studied than their state components (Grossnickle, 2016). Well established, generally-applicable, and well-verified measures of state curiosity could not be located in the literature.
An extensively studied area within games and training is presence and immersion (Coelho et al., 2006; McMahan et al., 2012; Nacke and Lindley, 2010). There are multiple components of presence and immersion. Spatial presence or sensory immersion is feeling as though one is in the virtual space (Nacke and Lindley, 2010). Narrative presence or imaginative immersion is being immersed in the story and the characters (Adams, 2009; Chen et al., 2011; Przybylski, Rigby, and Ryan, 2010; Rigby and Ryan, 2007; Ryan, Rigby, and Przybylski, 2006). Emotional presence is feeling that game events have real emotional weight. Finally, challenge-based immersion is similar to flow: Being so absorbed in the play of the game that one forgets about the external world (Adams, 2009; Nacke and Lindley, 2010). Such tools may be appropriate if the theories behind one’s serious game revolve around the motivational power of immersion to achieve the desired outcome. For abstract games that have no real backstory, world, or characters, immersion measures are likely inappropriate.

The two most popular measures in games research are the Player Experience of Need Satisfaction (PENS, Rigby and Ryan, 2007), and the Game Experience Questionnaire (GEQ, Law, Brühlmann, and Mekler, 2018). There is some potential for confusion because there is another GEQ: The Game Engagement Questionnaire (Brockmyer et al., 2009). For the sake of clarity, the former will be termed GExpQ, and the latter GEngQ. In a comparative review of the two, Norman (2013b) concluded that both were sub-optimal tools for evaluating the success of a game. The GEngQ was developed specifically to study people’s tendency for gaming as a stable character trait for the purposes of investigating the effects of videogame violence (Brockmyer et al., 2009). The GExpQ was developed to evaluate the fun of playing a game, but was built in a bottom-up manner from group interviews with players and has suffered from problems of factor structure and unreliability (Brühlmann and Schmid, 2015; Johnson et al., 2014). In contrast, PENS was built in a top-down manner from Self-Determination Theory in motivational psychology (Deci and Ryan, 2000), which posits that intrinsic motivation arises from needs for autonomy (e.g. agency and player choice), competence (e.g. success, victory, and high scores), and relatedness (e.g. competitive or cooperative play with friends). In a comparison of the validity of the GExpQ and PENS, Brühlmann and Schmid (2015) and Johnson
et al. (2014) found PENS generally superior in terms of consistent factor structure and reliability (see also Klarkowski et al., 2015). Similar results were found again by Johnson et al. (2018). Due to its being the more rigorously analysed and having foundations in established psychological theory, PENS seems at this stage a justifiable choice of motivation measure to help triangulate engagement.

Time spent playing is a direct measure of the duration component of engagement defined above. It is often treated as an indicator of liking a game (Johnson et al., 2014; Malone, 1980). If it is time that is voluntarily spent on the game, then it is a plausible indicator of intrinsic motivation to play. For example, in many studies on intrinsic motivation, free time spent on an activity is often used as a measure of intrinsic motivation (Boggiano and Ruble, 1982; Cameron, Banko, and Pierce, 2001; Elliot and Harackiewicz, 1996; Lieberoth, 2015; Ryan and Deci, 2000). To ensure that participants are spending time on the activity out of enjoyment and not just to avoid suffering stimulus deprivation, studies must include alternative activities that participants could choose to pursue during the free time period. Common alternative activities in the literature include magazines (Elliot and Harackiewicz, 1996) and games or toys such as puzzles, marbles, mazes, and art materials (Boggiano and Ruble, 1982; Elliot and Harackiewicz, 1996). Free time spent could also be compared to the time required for learning to test the efficiency of an educational game design. This makes time a very useful variable in this field of research.

### 2.2.4 Play

Extensive research has failed to produce consensus on what exactly play is and why people do it (Bateman, 2014; Boyle et al., 2012; Hamari and Tuunanen, 2014; Mellou, 1994; Rakoczy, Warneken, and Tomasello, 2008; Rubin, 2001b; Smith and Vollstedt, 1985; Steen and Owens, 2001; Turnbull and Jenvey, 2006). For example, one perplexing finding is that Ravaja et al. (2005; 2006) and Salminen and Ravaja (2008) measured indications of positive affect following from failure events in a game, and Abbott (2011) observed a player intentionally failing the game they had set up (such findings motivated study one of this thesis). The two questions of ‘what’ and ‘why’
are heavily interdependent, because there are many ways to play. Play involves a wide variety of behaviours that can each be motivated by different things. For example, play as a social bonding activity is very different from play as a form of exploration and experimentation.

All of these theories and studies defined play in different ways. These differences in definitions were because they were interested in a specific context or question. But the myriad definitions do have some things in common:

- **Intrinsically motivated**: Play is pursued for its own sake, not due to an external reward or punishment.
  - **Fun**: Play is enjoyable.
  - **Voluntary**: Play is something that is freely chosen, not mandated externally.

- **Activity**: Play requires taking action, or interacting with something – one cannot play by lying back and doing nothing. That is just resting or viewing or listening. In order to play, you need to participate.

These seem to be the minimum requirements to include when trying to understand play in some context. Additional elements are often added depending on the topic under investigation. But this does not tell us what play is or why people do it. Theories of play motivation may offer insight on that question, which will be particularly relevant to RQ2 concerning how play motivations could conflict or interact.

### 2.2.4.1 Challenge

The concept of ‘challenge’ is inescapable in the game design literature when discussing play motivation. Where one finds game design, one finds challenge. Empirical research has confirmed that players play to be challenged (Alexander, Sear, and Oikonomou, 2013; Boyle et al., 2012; Hoogen et al., 2012). The mastery hypothesis (Boggiano and Ruble, 1982) and effectance motivation (White, 1959) posit that people
are motivated to do something specifically because they find it difficult, and so de-
velop their skill at the task. Self-Determination Theory predicts that people will be
motivated to affirm their competence through undertaking challenging tasks. This
theory has received experimental support in the domain of games (Deci and Ryan,
2000; Ryan and Deci, 2000; Rigby and Ryan, 2007; Przybylski, Rigby, and Ryan,
2010). Players report challenge as a more important reason for play than multiple
other motivations (Sherry and Lucas, 2004).

The psychological theory of flow states that people will experience enjoy-
ment when their skill is matched with the challenge of the task (Czikszentmihalyi,
1990), although it includes many other elements that are not necessarily related
to challenge levels. This theory has experimental support (see e.g. Sweetser and
Wyeth, 2005; Engeser and Rheinberg, 2008; Klarkowski et al., 2015). It is often
stressed that the level of challenge in a game must not be so high as to be frustrat-
ing, but also not so low as to be boring (Przybylski, Rigby, and Ryan, 2010; Nacke
and Lindley, 2010). There needs to be a medium level so that the game is challeng-
ing in an enjoyable way (figure 2.3). This is one of the most basic ideas presented by
Czikszentmihalyi (1990) as part of his larger, more complex theory of flow to explain
optimal experience during an activity. Thus, this prediction of an ideal intermediate
challenge level is convenient to refer to as a flow effect (as in e.g. Klarkowski et al.,
2015), even though it is a small part of the theory, and so failure to find support for
the prediction would not actually disprove the theory of flow. This “sweet spot” of
intermediate challenge will become very relevant to the later discussion on complex-
ity theory. Included in this conception of challenge is the importance of mastering
the game system. After all, as pointed out by Togelius and Schmidhuber (2008), a
game is not fun if it can be won by acting randomly.
The common theme in play motivation (e.g. Flow theory) is that the challenge cannot be too great or too small, otherwise the player will be either frustrated or bored. This inverted U-shape of graph will be continually relevant throughout this literature review, especially with regard to complexity theory.

### 2.2.4.2 Juice

Play could also be motivated by the simple satisfaction of acting on the world to produce a reaction. Studies have found children play with objects specifically for the interesting sounds and reactions they produce (Ruff, 1984; Rochat, 1989). Several theories of play suggest it is a means of maintaining an optimal level of sensory stimulation and cognitive arousal (see e.g. White, 1959; Keller, Schneider, and Henderson, 1994). Effectance is the pleasure in affecting one’s environment (in other words, the drive to do something because it does something. See White, 1959; Klimmt, Hartmann, and Frey, 2007; Rogers, Dillman Carpentier, and Barnard, 2016).

In game design literature, this kind of play motivation is often called *juice*, *polish*, or *theatrics* (Schell, 2008; Adams, 2009; Jonasson and Purho, 2012; Brown, 2015; Swink, 2009; Stout, 2010; Tornqvist and Tichon, 2019). This refers to the feedback from the game as the player interacts, pleasing the senses with spectacular sights and sounds. Hicks et al. developed a framework for juicy game design (Hicks et al., 2018). The concept of juice emphasises the importance of sensory satisfaction from interaction. Gray et al. (2005) describe it as
Constant and bountiful user feedback... A juicy game feels alive and responds to everything you do — tons of cascading action and response for minimal user input. It makes the player feel powerful and in control of the world... (p. 3)

As such, juice can be an integral part of the moment-to-moment interactions and mechanics of a game. This concept from the game design literature is plausible considering extensive psychological studies on the effect on motivation from different kinds of feedback or rewards (e.g. Boggiano and Ruble, 1982; Cameron, Banko, and Pierce, 2001; Deci and Ryan, 2000; Hoska, 1993). They can be differentiated into categories such as tangible or verbal, and information or reward. But whether juice is necessarily a form of verbal information feedback or extrinsic reward is debatable. The gratuitous spectacle of juice is not strictly necessary or informative (no more so than a dry, technical form of informational feedback), and juice is often intended as a reward so in that sense perhaps it is an extrinsic incentive. In which case, the reward from a specific juicy effect could diminish with repeated exposure. But juice also isn’t tangible and has no extrinsic value outside of the game, which arguably makes it a kind of endogenous, informational feedback. One could argue that juice simply makes the information more salient, and therefore is best considered just as an element of information communication. However, the above definitions of juice describe its capacity to motivate outside of mere communication. For example, a catastrophic death of the player character is a negative communicative cue indicating that the player should not do that again, but it is also very juicy, which could encourage the player to do it again. In that case the communicative component and juice motivation component seem to conflict. This would suggest juice is not simply informational but has an independent motivational component.

To say that juice is merely a communicative tool, or an extrinsic incentive, would dismiss play purely in pursuit of juice as “not real play”, because play is intrinsically motivated. Intrinsic motivation is typically characterised as something done voluntarily and for its own sake. The fact that people spontaneously choose to pop bubble wrap and knock over sand castles would imply that they are intrinsically motivated to pursue these actions. However, juice could be added to an otherwise
dull activity to try to motivate behaviour and in that sense possibly is an extrinsic reward. It is unclear whether juice is likely to be perceived as a “controlling” reward, or merely an “affirmation of competence” for performing an action (Deci, Koestner, and Ryan, 1999). Future work will need to investigate that distinction. For the purposes of this study, one can remain agnostic as to its intrinsic or extrinsic nature, and simply call it a motivator. That is how I will proceed.

*Effectance* is the pleasure in affecting one’s environment (in other words, the drive to do something because it does something. See Klimmt, Hartmann, and Frey, 2007; Rogers, Dillman Carpentier, and Barnard, 2016; White, 1959). Several theories of play suggest it is a means of maintaining an optimal level of sensory stimulation and cognitive arousal (see e.g. Keller, Schneider, and Henderson, 1994; White, 1959). For example, some people choose to receive an electric shock rather than sit and do nothing, seemingly preferring to experience a negative stimulus rather than no stimulus at all (Wilson et al., 2014). Juul and Begy (2016) tested the effect of juice on outcomes such as perceived game quality and personal competence, but found only small effects that were not statistically significant. No studies could be located directly testing the capacity for juicy feedback to motivate behaviour such as interfering with the pursuit of a challenging goal. Study one, presented in this thesis (page 117), is the first to directly compare juice and challenge.

From the above discussions of juice, it is clear that it is not just informational feedback. In fact, it is specifically unnecessary feedback (Juul and Begy, 2016). The necessary information (e.g. confirmation of a hit) is present, but it is not strictly utilitarian. Juice is the extent to which this informational feedback is embellished, exaggerated, or decorated with gratuitous, non-informative elements. Therefore, any studies of juice would have to compare a ‘non-juicy’, strictly informational form of feedback to a ‘juicy’ form of feedback that presented no more or less relevant information, but included unnecessary additional elements.

The above definitions of juice by game designers are not scientifically rigorous. But if scientific support for this broad conjecture by game designers is found, then it would be worthwhile to develop the juice concept into a formal model that
yields much more precise predictions. But the first question to answer is if juice has any validity at all (for this, see study one of this thesis, and recent studies such as Hicks et al., 2019). To investigate this fledgling concept, juice can be defined as the perceptual magnitude of feedback from an action. This ‘magnitude’ is likely composed of various dimensions such as the salience (e.g. louder, brighter, bigger), complexity (e.g. detail, variability, predictability) and malleability (i.e. how much of the player’s input is reflected in the output – is the effect pre-determined, or shaped entirely by the player, or somewhere in between?). There are also additional questions to resolve such as how best to operationalise these dimensions (e.g. using information theory measures of predictability in stimuli, or physiological measures of affect and arousal). However, at this stage there is insufficient scientific data from which to form such a comprehensive explanatory and predictive theory of juice. Before elaborating all its nuances, one needs to first establish if it is even reproducible in experimental conditions (see study one, page 117).

### 2.2.4.3 Game Value Adoption

There is a popular assumption that players need or want to be told what to do (e.g. Adams, 2009; Garris, Ahlers, and Driskell, 2002; Gee, 2005a; Juul, 2005; Portnow and Floyd, 2014; Salen and Zimmerman, 2003). Supposedly players voluntarily enter a contract of arbitrary game rules, with the expectation that those rules would lead to enjoyment. From this perspective players are eager to be told how to have fun, evidenced by their decision to play the game in the first place. Adams (2009) explains, "The player tries to read the designer’s mind to some extent, to figure out what you want him to do, and then he does it" (p. 608). The importance of having clearly communicated goals is another strong thread in the game design literature (e.g. Adams, 2009; Gee, 2005a; Malone, 1980; Schell, 2008).

Text could directly instruct players on what to do. A scoreboard could provide real-time feedback on what actions are good or bad by gaining or losing points. Visual symbolism could be used as a subtler form of communication. Or a fully realised fantasy setting and narrative could establish heroes, villains, obstacles, etc.
The above authors posit that, however it might be communicated, once a player is given an impression of what is "good" or "bad" according to the game (what the game values or de-values), the player tends to adopt these same values as best they are able to interpret them, in order to play the game. This concept could be called ‘arbitrary goal compliance’, but here, I will use the term ‘game value adoption’ in order to incorporate a plausible mechanism by which it operates. Of course, players do not always adopt the game’s values (they might deliberately or accidentally transgress during play), but nor do all players always play for challenge or juice. Each concept of play accounts for one of many possible ways to play.

The game value adoption concept proposes one particular mechanism influencing play: players adopt the values and priorities of a game’s rules or world, as they are able to interpret them based on perceptible cues or feedback. It proposes an implicit contract that players voluntarily enter in choosing to play a game – to follow the rules to have fun. This is arguably similar to a player’s willingness to pursue arbitrary badges or achievements just because they are there (see e.g. Hamari, 2017). To adopt whatever the game defines as ‘good’ or ‘bad’, to adopt its values and priorities, is a fascinating and wilful act. Adams (2009) explains:

At first glance, you might not think much pretending is involved in a physical game like soccer. After all, the players aren’t pretending to be someone else, and their actions are real-world actions. Even so, the players assign artificial significance to the situations and events in the game, and this is an act of pretending. (p. 5)

This may seem like a trivial concept - players try to win games - but it is worth discussing how and why clearly communicated goals could motivate, and therefore how they could fit as a variable in a larger conception of play. In particular, it is worth noting how goals function differently in game value adoption when compared to other theories of play that involve goals. When the valued state (the goal) is derived from the fictional setting of the game, there is some overlap into the territory of immersion and suspension of disbelief. Therefore, one could mistakenly refer to the theory of play, fantasy (Malone, 1980), as being equivalent to game value adoption.
adoption. However, note that the fantasy play theory merely predicts that games will be more motivating with an exciting or interesting fictional presentation layer. In contrast, game value adoption does not require the presence or absence of a fictional presentation layer. Even if the game is purely abstract, game value adoption predicts players will be motivated to achieve the goal. A goal can be communicated with a fantasy presentation layer, but it doesn’t have to be.

Similarly, a goal can be (but need not be) communicated by feedback. For example, initially a player might have no idea what they are meant to do, but when they hit an object, their high score increases by ten points. From this, one might mistakenly conclude that game value adoption is equivalent to established psychological theories of informational and positive feedback. However, just as with fantasy, feedback is merely one possible means of communication, and is not strictly required according to game value adoption. Even if the player receives no feedback on completing the goal, game value adoption still predicts they would be motivated to achieve that goal. Therefore, its predictions are divergent from these other theories because game value adoption makes the fewest assumptions or requirements on how the goal is communicated to the player. But it does propose an interesting psychological mechanism to explain that motivation: players adopt the values of the game’s rules or world.

The voluntary acceptance of arbitrary constraints or goals is common in attempts by psychologists to classify play from field observations (e.g. Rakoczy, Warneken, and Tomasello, 2008; Rubin, 2001b; Smith and Vollstedt, 1985; Turnbull and Jenvey, 2006). Malone (1980) found experimental support for the idea that games with goals are more enjoyable than those without, and Lieberoth (2015) found that players enjoyed an activity more by merely framing it as a game, all of which give some plausibility to game value adoption. But, despite its prominence in game design writings, study one of this thesis will be the first study to test the central prediction of game value adoption.

Game value adoption is an under-developed concept that raises additional questions that future studies could explore to elaborate it into a theory. For example,
if it is true that the reason why communicated goals are motivating is that players adopt the values of the game, then what are the minimum conditions for a player to accept that something is a game, and therefore adopt its values and pursue arbitrary goals? People aren’t generally receptive of arbitrary goals if they are just shouted out while walking in the street, so what conditions are necessary to make the context sufficiently gamelike for people to adopt these arbitrary values? Is there a point at which those values are so extremely different from the players’ own values that they cannot accept them (for example, if a game asks the player to kill innocent virtual characters)? Does this acceptable difference in values scale linearly with making the game more gamelike (i.e. not just the minimum conditions for establishing a game are met, but taken further to reinforce the idea that this is a game)? In this way it may become an additional factor in the moral buffer (Cummings, 2004), which seeks to explain humans’ willingness to launch missiles from a console in terms of factors that distance oneself from the recipient of the decision to kill. In that context, there is passing mention of gamelike features, but they are neglected in favour of factors such as automation and remoteness. Additional research on gamelike cues could be invaluable. Game value adoption is in need of much further study and elaboration, but this is beyond the scope of this research.

Clearly the depth of the topic of play is vast. Rather than probing people’s motivations to play, their psychological needs, or their subjective experiences of enjoyment, this research will simply be concerned with whether or not people play. I will not be classifying a comprehensive list of different play behaviours, but limit my focus to one or more specified forms of play at a time. For example, social play is not relevant to this study: I will not be expecting it to occur, I will not be trying to provoke it, and I will not be measuring it. The possibility of multiple forms of play coming into conflict forms RQ2: How can different combinations of game features interact to cause players to engage in off-task behaviour instead of trying to win the game (and thereby likely miss the purpose of an educational or serious game)?

But for the other two research questions (RQ1 "How can serious games research move from proof-of-concept studies towards building up a formal model of how specific design changes affect specific outcomes such as learning and engage-
ment?” and RQ3 “What is the relationship between the complexity of a game, and
the player’s ability to master it?”) the most pertinent form of play to focus on is
the kind that might explain how players teach themselves complex systems: Ex-
ploratory play.

2.2.5 Exploratory Play

The high-level cognitive theory underpinning this research is that of exploratory
play (Tornqvist, 2014).

The dominant paradigm for educational games has been constructivism,
wherein learning is regarded as a process of knowledge construction by the stu-
dent, possibly with the guidance of a teacher (Badilla-saxe, 2010; Baytak, Land,
and Smith, 2011; Liu and Matthews, 2005; Mitra and Dangwal, 2010; Papert, 1980;
Resnick, 2006). It is contrasted with the traditional instructional paradigm wherein
the teacher hands down knowledge to the passive student. Instead constructivism
emphasizes exploration, discovery and individualized approaches, making it a natu-
ral fit for game-based learning. The capacity for spontaneous exploratory play to de-
velop one’s understanding has been observed in constructivist learning paradigms
(e.g. Badilla-saxe, 2010; Baytak, Land, and Smith, 2011; Mitra and Dangwal, 2010;
Papert, 1980; Resnick, 2006), and has received experimental support (e.g. Cook,
Goodman, and Schulz, 2011; Gopnik and Schulz, 2007; Jennings et al., 1979; Schulz,
Standing, and Bonawitz, 2008).

But there has also been counter-evidence produced (e.g. Doyle, Radzicki,
and Trees, 1998; Frank, 2011; Mayer, 2004). In particular, in system dynamics exper-
iments participants are notoriously bad at trying to control even relatively simple
dynamic systems (e.g. Jensen and Brehmer, 2003). In a review of the literature,
Mayer (2004) concluded that learning by pure unguided discovery had been blindly
over-praised in comparison to the evidence of efficacy. This implies not that ex-
ploratory play is useless, but that it has been applied in contexts where it was not
appropriate, or should have occurred alongside guidance or instruction. Bonawitz
et al. (2011) agree that the efficiency of direct instruction cannot be denied, but they
found that it came at the cost of exploration. They conclude that direct instruction is well suited for contexts where the teacher has the time and resources to impart all the needed knowledge, but if they do not then it might be best to maximize the tendency to explore instead. “Videogame syndrome” refers to mindless reaction to system feedback without developing one’s mental model of the system (Sterman, 2000).

However, Cook (2006; 2007) and Koster (2005a) would likely regard “videogame syndrome” as a misnomer, because they argue that the fun in games is derived entirely from the process of learning how to interact with and master the game system. For example, Self-Determination Theory’s (Deci and Ryan, 2000; Ryan, Rigby, and Przybylski, 2006) emphasis on the psychological need for autonomy and competence correspond to exploratory play’s emphasis on open-ended choice (autonomy) and learning (developing competence). Along with the work of Gopnik (2011), Loewenstein (1994), Schulz et al (2008), Sutton-Smith (1975), Ruff and Saltarelli (1993), Weisler and McCall (1976), and White (1959), these strongly imply that designing an inherently fun system is equivalent to designing an inherently learnable system. This concept of an inherently learnable system received much incidental and tangential discussion, but little direct analysis or testing. Addressing RQ3 and RQ1 required gathering and consolidating the literature on this concept to assess what evidence there was that it could occur, under what conditions, and for what purpose. It was concluded that inherently learnable systems epitomise a paradigm of education of immense importance that, much like inherently learnable systems themselves, also has received little explicit acknowledgement or investigation. In a paper currently under review I term this paradigm Massively Scalable Learning (MSL) and outline its major components, mechanisms, and appropriate domains of application, as well as what is currently known about how to design for massively scalable learning, and where the gaps in knowledge on this topic remain. The full details of this contribution can be found in appendix A.

Because motivation is an important element of learning, it would be beneficial to look at this claim about game systems, and discover whether the systems that are fun to play with are also supportive of developing understanding and mastery.
If so, this would offer one solution to the “chocolate on broccoli” problem. Therefore, the principle aim of this research is to improve our knowledge of how to design interactive systems that are inherently learnable by both provoking exploratory play and facilitating the development of understanding and mastery of its systems.

Although seeming like nothing more than spontaneous fun to the player, exploratory play is motivated by curiosity and the drive for mastery. Experiments have demonstrated that unguided children will teach themselves how novel objects work: When playing they intervene to resolve ambiguity, and they retain this knowledge of causal function to use later in achieving their goals (Cook, Goodman, and Schulz, 2011; Gopnik and Schull, 2007; Jennings et al., 1979; Schulz, Standing, and Bonawitz, 2008).

Curiosity is considered one of the primary motivations of exploratory play (Keller, Schneider, and Henderson, 1994; Loewenstein, 1994; Schulz, Bonawitz, and Griffiths, 2007; White, 1959; Lorenz, 1981). Curiosity is considered both a trait and a state (Naylor, 1981), meaning that curiosity is both a stable characteristic that can vary across individuals, and a state of mind that can vary over time within an individual. As a character trait, curiosity could be a central causal variable of the player in the process of playing and mastering complex games, as more curious individuals may be more intrigued by complexity in games, or more disposed to exploratory play.

Curiosity can also be separated into whether it is specific or diversive, and whether it is epistemic or perceptual (Naylor, 1981; Reio et al., 2006). Specific curiosity is a focus on discovering a particular piece of information. At the other end of the spectrum, diversive curiosity is concerned with learning something that is not specified in advance. It can also be epistemic or perceptual. Epistemic curiosity is concerned with learning information. Perceptual curiosity is concerned with having a novel experience. Therefore, since exploratory play involves doing something just to see what it does, it can be categorised as a form of diversive epistemic curiosity.

There are many theories in psychology that attempt to explain curiosity, but there are several broad categories of these theories that are relevant to this research.
Drive theories posit that it is an internally-arising need (similar to hunger or thirst) that compels information-foraging or stimulation-seeking behaviour (Berlyne, 1954; Grossnickle, 2016). This explains why people may seek out puzzles, games, mysteries, and skills to learn of their own volition. In contrast, incongruity and knowledge-gap theories posit that an external stimulus provokes curiosity by revealing a conflict with one’s conception of reality, or revealing a gap in one’s understanding of reality (Grossnickle, 2016; Loewenstein, 1994). This explains why curiosity can be provoked by tantalising partial information, or a surprising and counter-intuitive event. Optimal-arousal theories posit that curiosity is a manifestation of needing to maintain a particular, ideal level of arousal, and that therefore people will seek novel stimuli when under-aroused, and recoil from the unfamiliar when over-aroused (Grossnickle, 2016). These different theories of curiosity have different predictions for its role in playing with complex games (see study two, page 166).

This form of play is most common in the simulation sandbox (sim sandbox) genre of game (Breslin, 2009; Calhoun, 2010; Francis, 2006; Squire et al., 2004; Stewart, 2014). This is due to many of the genre’s characteristics fostering curiosity and removing barriers to spontaneous experimentation (Tornqvist, 2014). For example, the sandbox component of the genre entails providing the player with an open-ended game system. Instead of steering them narrowly toward one goal, a sandbox empowers the player to explore the possibility space of the game’s systems at their own whim. The absence of a specific goal is a critical condition necessary for learning causation in research on solving problems in complex systems (Geddes and Stevenson, 1997; Kistner et al., 2015; Künsting, Wirth, and Paas, 2011; Wellen, 2014; Wirth, Künsting, and Leutner, 2009).

Another central characteristic of sim sandboxes is their simulation component. Since many games occur in fictional contexts, here “simulation” does not mean “representing a real phenomenon”, but merely an in-depth systemic representation of some kind of phenomenon, real or not. Most game genres include only very basic approximations of phenomena insofar as it adds to the challenge of completing the goal. But sim sandboxes, lacking a goal, provide very elaborate simulations of the
phenomena they depict. This complexity has the potential to motivate exploratory play through emergent gameplay.

Emergent gameplay is frequently discussed as a core design feature in games that promote exploratory play (e.g. Kickmeier-rust and Albert, 2009; Sweetser, 2008). A common definition of emergence is a simple set of rules that give rise to complex overall system behaviour (e.g. Crutchfield, 2011; Jost, Bertschinger, and Olbrich, 2010; Garneau, 2008a; Sweetser, 2008). Unexpected phenomena pique one’s curiosity (Berlyne, 1954; Berlyne, 1966; Day, 1982; Loewenstein, 1994), which is a central motivation of exploratory play. For a game designer aiming for emergent gameplay, the aim is for the game to behave in ways that surprise even the designer. The system is of such complexity that its full implications cannot be intuited even from complete knowledge of its structure.

The idea of emergent gameplay implies that developing explicit knowledge of the underlying rules would be independent of implicit mastery of the patterns of system behaviour, and vice versa. The literature on complexity theory provides interesting mathematical tools and concepts relevant to the primary research problem of developing a framework for educational game design, and RQ3 concerning the relationship between the complexity of a game and the player’s ability to master it.

2.3 Complexity

Complexity is a critical variable in the primary problem of a unifying framework for serious games research, and RQ3 concerning how complexity relates to curiosity in games. One of the most striking and consistent findings is people’s difficulty at solving problems within relatively simple dynamic systems (Meadows and Wright, 2008; Rouwette, Größler, and Vennix, 2004; Sterman, 2000). “Stock and flow failure” is the term for the pervasive inability to infer the behaviour of dynamic systems (Fischer and Gonzalez 2015). This has highlighted the need for research into improving people’s understanding of complex systems. Despite this, studies rarely quantify complexity using tools from complexity theory.
Complexity has been defined as the amount of unpredictability in a stream of information: Kolmogorov Complexity (Prokopenko, Boschetti, and Ryan, 2008). But there are myriad other definitions, such as the balance between change and stability (Fernandez, Maldonado, and Gershenson, 2014), self-dissimilarity at different scales (Wolpert and Macready, 2004; Wolpert and Macready, 2000), among others (for reviews, see Bonchev and Buck, 2005; Chu, Strand, and Fjelland, 2003). With no consensus on which one is “correct”, it is left to the discretion of individual researchers to select (or develop) a complexity measure that is most suited to answering their research question.

One possible reason measures produced by complexity theorists are rarely used in cognitive science, is that they generally remain too abstract, concerned more with technical discussion of the mathematical essence of complexity than with predicting human cognition. For example, Emmert-Streib (2010) and Gershenson and Fernandez (2012) tested their models by applying them to data generated by a software algorithm. A conception of complexity more relevant to human cognition might be found by considering a companion concept: Emergence.

2.3.1 Emergence

Emergence has proven just as troublesome as complexity for cognitive scientists and educators. Teaching students about emergent processes has proven difficult, as misconceptions have proven tenacious (Charles and Apollonia, 2004; Chi, 2005; Chi et al., 2012; Jacobson et al., 2011; Wilensky and Novak, 2010).

There is much debate around the exact definition and philosophy of emergence (e.g. Atay, 2011; Corning, 2002; Chalmers, 2002; Johnson, 2006; De Wolf and Holvoet, 2005; Deguet, Demazeau, and Magnin, 2006; Goldstein, 1999; Jost, Bertschinger, and Olbrich, 2010; Kim, 1999; O’Connor, 1994; Pepper, 1926). For example, Goldstein (1999) defines emergence as, “the arising of novel and coherent structures, patterns, and properties during the process of self-organization in complex systems”. Depending on the author, emergence may or may not entail the following properties:
• Radical novelty, surprise, unpredictability and inexplicability: An emergent phenomenon is novel even with full knowledge of its constituent parts and the rules by which they interact. According to some, it is not just surprising, it is not even reducible to or deducible from its constituent structure (Buchmann, 2001; Chalmers, 2002; De Wolf and Holvoet, 2005; Henle, 2009; Hosseinie and Mahzoon, 2011).

• Coherence, stability, resilience: An emergent phenomenon is not a transient, fleeting moment. It is a stable and coherent phenomenon, and according to some, even resilient, adaptive and self-preserving (De Wolf and Holvoet, 2005; Fromm, 2005; Moncion, Amar, and Hutzler, 2010).

• Discontinuity between micro and macro scales: There are (at least) two distinct scales, which operate (or appear to operate) according to different rules. These can be physical scales (e.g. the quantum scale vs. the molecular scale), or time scales (e.g. a matter of nanoseconds, or a matter of centuries).

• Synergistic and heteropathic effects: The whole is more than the sum, in fact it is qualitatively different from the sum of the parts. The classic example is table salt: sodium chloride. Individually, either element would be dangerous to your health, but together they form something that your body needs to operate healthily.

• Downward causation: The rules operating at the macro scale feed back into the micro scale to affect it. For example, the placebo effect is arguably a case of the mind affecting the physical body chemistry (Brodu, 2008).

Each of the above properties only provide a qualitative and binary assessment of whether or not something is emergent, and none have been agreed upon as either necessary or sufficient for emergence. This makes it difficult for anyone trying to determine if a specific phenomenon or system is emergent.

Some have proposed a distinction between weak and strong emergence (Chalmers, 2002). A phenomenon is weakly emergent if it is ultimately deducible from knowing the underlying structure, but is merely unexpected. A phenomenon
is strongly emergent if it is not deducible from knowing the underlying structure. Brodu (2008) and Jost et al. (2010) have gone further and suggested that emergence should not be conceptualised as a number of discrete categories, but a relative measure along a continuous spectrum of emergence – a thing is only emergent to a degree, relative to another thing that is less emergent.

Philosophers and scientists still debate (Aksentijevic and Gibson, 2012; Boschetti and Gray, 2013; Brodu, 2008; Deguet, Demazeau, and Magnin, 2006; Goldstein, 1999; Ryan, 2007) whether emergence is best considered:

A an objective phenomenon (ontological emergence), falling within the domain of complexity theory or information theory; or

B a subjective phenomenon (epistemic emergence) that depends on the properties of the observer, in which case cognitive psychology may be a better tool to study emergence.

The search for a coherent and precise formal definition of emergence has focused on finding the conditions in complex systems where a different set of rules applies – rules for the emergent phenomenon. However, this assumes emergence is objective. Emergence can be a useful concept when considered a subjective phenomenon dependent on an observer. Taking this perspective will prove useful in this research project when considering how conceptions of complexity and cognition can interact to produce the experience of emergence within complex systems.

The conceptualisations of emergence are many, but the idea of simple rules causing complex behaviour is a very common theme among them (Abbott, 2006; Atay, 2011; Berrondo and Sandoval, 2015; Boschetti and Gray, 2013; Crutchfield, 2011; Nunn, 2007). That will therefore be the working definition from here on. Such a theory of epistemic emergence maps almost directly on to dual process theories of cognition that split implicit from explicit cognition: Learning about the simple rules (explicit knowledge) would not be sufficient to predict the complex behaviour (implicit knowledge), or having learned the complex behaviour, it would be difficult to infer the simple underlying rules.
This common definition of emergence implies that developing explicit knowledge of the underlying rules would be independent of implicit mastery of the patterns of system behaviour, and vice versa. If emergent gameplay is an effective design for generating interest in play, then according to this definition of emergence, it would do so by separating implicit and explicit learning (as discussed by Evans, 2012; Kahneman and Klein, 2009; Osman, 2004; Smith and DeCoster, 2000; Sun and Zhang, 2004). This dichotomy of parallel learning processes has been observed in cognitive experiments, forming the basis of dual process theories (e.g. Harteis et al., 2012; Sloman, 1996; Sun et al., 2007). Therefore, the definition of emergence implies that explicit learning of rules and implicit mastery of behaviours should only separate into independent learning processes when there is sufficient complexity – and that explicit and implicit learning can be reunited into the one process at lower levels of complexity. This implied prediction has yet to be tested. It describes a specific relationship between variables of complexity and learning, which can be tested empirically. Investigating this definition of emergence requires considering how the complexity of the rules (static or structural complexity) could be measured independent of the complexity of the behaviour (dynamic complexity).

2.3.2 Different Kinds of Complexity: Static vs. Dynamic

The common definition of emergence as simple rules giving rise to complex behaviour, implies a distinction between the complexity of the rules versus the behaviour. I will use the terminology static complexity and dynamic complexity to distinguish the two. Static complexity refers to conceptions and measures of complexity based on the fundamental elements and/or structure of the system (e.g. its rules, data structure, or number of components), whereas dynamic complexity refers to conceptions and measures of complexity based on tracking changes that develop over time as the system is active (e.g. its behaviour, patterns of output, or its response to different conditions or interventions).

An example of dynamic complexity can be found in chaos theory. Emergence has been described as occurring at the edge of chaos (Adams and Dormans,
2012; Brodu, 2008; Chalmers, 2002; Chi et al., 2012; Fromm, 2006; Hudson, 2011; Prokopenko, 2013; Sweetser, 2008). Too little chaos and the system is too ordered, making its behaviour simple and predictable. Too much chaos and the system is an incoherent mess, with behaviour that is unpredictable and thus indiscernible from random noise. At an intermediate level, at this hypothesised edge of chaos, emergence occurs and the system’s behaviour is complex without being simple or incoherent (figure 2.4).

The presence of certain properties like topological mixing and strange attractors are also signs of chaos (Mitchell, 2009; Sterman, 2000), but the central element to chaos is sensitivity to initial conditions (Gleick, 1988; Nunn, 2007). This can be measured along a continuous scale using the Lyapunov exponent (Cross, 2000; Habib and Ryne, 1995; Rosenstein, Collins, and De Luca, 1993). Finding the Lyapunov exponent of a system involves taking multiple nearby starting conditions, simulating them forward in time by a certain number of timesteps, and then measuring by how much they have diverged or converged (for detailed formulas, see Cross, 2000; Habib and Ryne, 1995; Rosenstein, Collins, and De Luca, 1993). A negative value indicates that they are converging (the initial difference between starting conditions is shrinking to make them more similar over time), a value of zero indicates they remain the same distance apart, and a value greater than zero indicates that they are diverging, indicating that the system is sensitive to initial conditions.
Formally, the Lyapunov exponent $\lambda$ is related to the rate of divergence of two trajectories in phase space with initial separation $\delta Z_0$ according to the following equation:

$$|\delta Z(t)| \approx e^{\lambda t} |\delta Z_0|$$  \hspace{1cm} (2.1)

The maximal Lyapunov exponent (as is generally sought when studying chaotic systems) can be defined as

$$\lambda = \lim_{t \to \infty} \lim_{|\delta Z_0| \to 0} \frac{1}{t} \ln \frac{|\delta Z(t)|}{|\delta Z_0|}$$  \hspace{1cm} (2.2)

where $\lim_{|\delta Z_0| \to 0}$ ensures the validity of the linear approximation at any time (for more technical details and methods to calculate it, see e.g. Cross, 2000; Habib and Ryne, 1995; Rosenstein, Collins, and De Luca, 1993. But a full understanding of how to calculate the Lyapunov exponent is not necessary to understand this thesis).

Network theory contains much discussion of static forms of complexity. Since game systems are often represented as a network of objects connected by lines of interaction (e.g. Adams and Dormans, 2012; Smith and Smith, 2004; Zupke, 2016), network theory is an apt domain from which to draw a theory of the complexity of game rules. Similar to the edge of chaos hypothesis, network complexity is considered highest at an intermediate level of connectivity – When there is too little connectivity, the network is considered relatively simple, but if there is maximal connectivity (i.e. every node is connected to every other node) then the network is also considered quite simple (Barabási, 2002; Bonchev and Buck, 2005). Some degree of connectivity between zero and complete connectivity is required to create a complex network (figure 2.5). The number of connections a node has to other nodes is termed its degree, and the probability distribution of these degrees over the whole degree is termed the degree distribution of the network. Therefore, the degree distribution is likely relevant to considering the complexity of a network.
Complex naturally-evolving networks (e.g. the internet) tend to be scale-free networks (Barabási, 2002; Mitchell, 2009; Wen, Kirk, and Dromey, 2007; Wen, Dromey, and Kirk, 2009), which are characterised in several ways (see previous citations for detailed mathematical descriptions of scale-free networks, but a full understanding of their formal description is not necessary to understand this thesis). Scale-free networks usually grow in a rich-get-richer manner whereby nodes with more connections are more likely to get yet more connections in future, concentrating connectivity among a relatively small number of major hubs, then a larger number of minor hubs with slightly less connectivity, and so on. Another attribute of these networks (as a consequence of the first) is that the degree distribution follows a power law (at least asymptotically). This pattern of degree distribution is commonly used as the formal mathematical definition of a scale-free network. This structure tends to make the network more robust to failure, in that the removal of a random node is unlikely to significantly affect the connectedness of the network, due to remaining hubs. Since the structure of scale-free networks is considered complex in network theory, it is worthwhile to investigate if such systems produce emergent behaviours.

Figure 2.5: Network theory suggests that the structural complexity of a network is greatest somewhere between complete connectivity, and zero connectivity. Specifically, when the structure forms a scale-free network.

Total walk count (TWC) is a measure of connectivity that involves counting the number of unique paths that can be traced between each pair of nodes. This
measure could be used to assess the level of inter-connectivity in a game’s systems, as a measure of a game’s static complexity.

Formally, TWC can be defined as

\[
TWC = \sum_{l=1}^{V-1} WC = \sum_{l=1}^{V-1} \sum_{i}^{l} w_i
\]  

(2.3)

where \( TWC \) is total walk count, \( WC \) is walk count, \( w \) is a walk, \( l \) is the length of a walk (TWC counting walks of all lengths), and \( V \) is the set of all nodes (vertices) in the network (details can be found in Bonchev and Buck, 2005, pp. 211-213. But a full understanding of how to calculate TWC is not necessary to understand this thesis).

Causal network connectivity and chaotic behaviour are elements of the game entity in the framework, and as such will form the independent variables of game complexity that I manipulate in my abstract game in study two.

Distinguishing these two forms of complexity enables direct relations to be drawn to dual process theories of learning in games - dynamic complexity may correspond to the difficulty of implicit learning of a game’s behaviour, and static complexity may correspond to the difficulty of explicit learning of a game’s rules. For a game to be emergent, the dynamics need to be difficult to infer from the statics and thus learned independently. Thus, leading to unexpected, emergent behaviours. Therefore, the definition of emergence used here (and implied in much discussion of emergence) is that the dynamics are difficult to infer based on knowledge of the rules - that they are each learned separately, not necessarily that they are different levels of complexity. This is the common theme among the majority of definitions of emergence, which one can pull into the domain of cognitive science to test its implications for designing inherently learnable systems. These concepts will prove invaluable when considering RQ3 of how the complexity of a game may relate to the player’s ability to master it.
All this diverse literature from complexity theory, game design, motivational psychology, and cognitive science, come together to provide a more comprehensive picture of how educational and serious games might achieve positive outcomes. Combining literature from these various fields proved crucial in addressing RQ1 (How can serious games research move from proof-of-concept studies towards building up a formal model of how specific design changes affect specific outcomes such as learning and engagement?) by developing the primary contribution of this thesis: a unifying framework to deconstruct the process into smaller sub-processes that can be studied individually, and to specify the broader relevance of narrower studies to allow them to be assembled to gradually build up a more precise and comprehensive model of how educational and serious games achieve positive outcomes.
Chapter 3

Proposing the Model-Master-Transfer Framework

The primary research question was: How can serious games research move from proof-of-concept studies towards building up a formal model of how specific design changes affect specific outcomes such as learning and engagement? To address this question, the primary contribution of this thesis is a formal framework to deconstruct serious games research into narrower sub-processes to be studied individually, and by which such narrow studies can be integrated into a larger mathematical structure.

Development of this new framework involved reviewing existing causal frameworks for educational video games (see literature review above, section Entities & Processes in Existing Frameworks, table 2.1) to identify the different major processes that are included and the major entities which are assigned a causal role in those processes. This review of existing frameworks informed the development of a new framework that incorporates these commonly-appearing entities and processes, and defines their relationships mathematically.

In reviewing the above models from the literature, I identified three major entities: the game, the player, and the real-world domain. Three processes were also identified: modelling the real-world domain into a game, the player master-
PROPOSING THE MODEL-MASTER-TRANSFER FRAMEWORK

ing the game, and then the player transferring that mastery back to the real-world domain. These entities and processes form the structure of my framework, Model-Master-Transfer (MMT). It provides a mathematical structure by which to study how these commonly-discussed entities and processes interact and relate to each other.

3.1 Defining Model-Master-Transfer

This framework considers educational video game research in terms of the three most commonly-appearing entities in table 2.1: The game, the player, and the real-world domain.

These entities are related via the three processes identified in the above review: The (video) game is modelled after the real-world task, the player masters the game, and then the player transfers that learning to the real-world task. This can be called the game-player-domain framework, referring to the entities, or the model-master-transfer framework, referring to the processes connecting the entities. Since the processes follow a logical, linear progression, I will use terminology that describes the processes connecting the entities (game, player, and domain). To that end, I refer to these processes and their interaction as the Model-Master-Transfer (MMT) framework. This framework incorporates the three major entities and major processes identified in the literature review, and provides a mathematical structure connecting them. Figure 3.1 illustrates the concept.
In future work many entities and processes could be added, or these major processes and entities could be subdivided into more specific sub-processes and sub-entities. For example, the player component could be further specified addressing aspects such as individual differences, background knowledge, attitudes, preferences, demographics. The MMT framework acts as a high-level umbrella concept that integrates these more detailed studies that focus on specific aspects. Consequently, MMT’s goal is not to replace existing concepts, but to provide a bridging structure by which these concepts can be compared for predictive power, or integrated to form a larger model that accounts for more processes. Therefore, MMT seeks to be parsimonious and agnostic towards many of the details of exactly how each of these three processes occur, and exactly which factors of these three entities are relevant.

Traditionally, most educational games studies have bundled together all three processes of modelling, mastery, and transfer (Barton et al., 2016; Cheng and Su, 2012; Hsu, Tsai, and Wang, 2012; Sengupta, Krinks, and Clark, 2015; Squire et al., 2004). This can be problematic when the game produced no measurable benefit to participants, and it is unclear where the failure occurred. As a result,
PROPOSING THE MODEL-MASTER-TRANSFER FRAMEWORK

the meta-analyses and literature reviews discussed above recommended research to shift away from these proof-of-concept studies and instead unpack these entities and relationships, and narrow the focus to investigating one at a time. The MMT framework provides a theoretical foundation for such studies.

In the following, I will focus on each process in detail (modelling, mastery, and transfer) and explain how mathematical formulae can precisely describe their interactions. I will use the example of educational video games being studied in terms of their complexity to elaborate these details of the MMT framework. By specifying mathematical relationships between these three processes, the MMT framework enables a study on just one process to inform our understanding and predictions of the other processes through these formulae. Thus, MMT is able to specify the broader relevance of narrower research.

This discussion will adopt the convention of using subscript to specify the object to which a variable belongs: \( g \) for the game, \( p \) for the person, and \( r \) for the real-world domain. The messiness and complexity of reality are temporarily set aside for the next section, **Mathematical Structure Overview**, in order to first present a precise formal mathematical explanation of the structure of MMT. This precision aids presenting a clear overview, but of course is very abstract. Therefore, the complexities of reality that were initially set aside are incorporated into the second section, **Usage of MMT**, to discuss them in much greater depth, and in more concrete rather than abstract terms, explaining MMT in relation to empirical experiments, measurable variables, and testable predictions.

### 3.2 Mathematical Structure Overview

*Definition 1: A model, denoted by \( m \), will describe a particular object, denoted by subscript (e.g. a model of a real-world phenomenon can be denoted as \( m_r \), and a model of a game can be denoted as \( m_g \)). The format of the model (whether it is, for example, a finite state machine, an undirected network, or other type of model) is*
unspecified by the framework, but it assumes that both models, $m_g$ and $m_r$, can be translated into comparable formats, without assuming what such formats might be.

**Definition 2:** **Veracity**, denoted by $V$, is a normalised variable describing the degree of correspondence between the two models, $m_g$ and $m_r$. There are many papers dedicated to the subject of how to derive a measure of the isomorphism of two systems (Schieber et al., 2017), and this framework does not specify the method to accomplish this, leaving it to the discretion of researchers to select or develop a veracity measure most appropriate to the subject matter and the model format. As such there can be various ways of calculating $V$, based on different conceptions of isomorphism and/or cognitive learning processes (e.g. IBLT suggests that the perceptual fidelity of the interface is a critical consideration in determining $V$). But the broad principle that the game model will contain a certain amount of information (or knowledge to be learned) that is abstracted and simplified from the real-world domain, can be written as:

$$0 \leq V = \frac{2K(m_g)K(m_r)}{K(m_g)^2K(m_r)^2} \leq 1$$  \hspace{1cm} (3.1)

Where $K(m)$ is the informational content of the model determined via Kolmogorov complexity (a proposal for how $K(m_g)$ could be calculated is presented in section Proposing Dynamic Causal Nets, below). To explain MMT, it is not necessary to say exactly how $V$ is determined, but this is one example, and various ways are discussed in a dedicated section below. Note that a $V$ of one would indicate that the two models are identical.

**Definition 3:** **Performance**, denoted by $\rho$, will be within a particular context (e.g. the high score achieved in the game, or the assessment of learning in a real-world context), denoted by subscript, as in $\rho_g$ to denote participant performance in the game, or $\rho_r$ to denote participant performance in the real world. As performance may be multi-dimensional, containing several measures of performance in different skills or knowledge topics, it is best considered a vector.
**Definition 4:** *Transfer*, denoted by $T$, is a normalised variable describing the amount of performance in the game, $\rho_g$, that is carried over into the real-world context, $\rho_r$, where a value of zero indicates that nothing has transferred, and value of one indicates that all learning in the game is transferred to the real-world domain. This can be written:

$$T = \frac{\rho_r}{\rho_g}$$  \hspace{1cm} (3.2)

**Axiom 1,**

$$T = V$$  \hspace{1cm} (3.3)

This would indicate a game with a high $V$, by virtue of having a high correspondence to the real world task, would allow players who have mastered the game to transfer a large amount of that mastery to the real world, resulting in a high $T$. Otherwise, a game with a low $V$ would have little correspondence to the real-world task and thus a low amount of transfer would be observed, resulting in a low $T$. The idea that $T = V$ may not be obvious but is an algebraic consequence of the following formulas that relate $T$ and $V$ to other variables.

When actual measures only allow approximating $T$ and/or $V$, then (3.3) will not hold, and it will instead be the case that $T \approx V$. Therefore, according to MMT, the difference between $T$ and $V$ indicates the margin of error in one’s methods of obtaining one or both values.

Formulas (3.2) and (3.3) are all that is needed to derive another formula that states that the relationship between performance in the game, $\rho_g$, and performance in the real world, $\rho_r$, is mediated by the veracity of the game (i.e., a game with more correspondence to the real-world domain it tried to teach would enable more direct transfer of mastery of the game to mastery of the real world), as in the equation:
This formula captures the notion that if a game that has no correspondence to the real-world domain \((V = 0)\), or if a participant has completely failed to master the game \((\rho_g = 0)\), then they will not be able to transfer any learning to the real-world domain \((\rho_r = 0)\). Conversely, if the game is a perfect simulation \((V = 1)\), then whatever level of mastery the participant has achieved, \(\rho_g\), should theoretically transfer perfectly to the real-world domain, \(\rho_g = \rho_r\). Although, of course, it would be virtually impossible to achieve a \(V\) of one.

**Definition 5:** **Play appeal factors**, denoted by \(a\), can be found within a particular entity (e.g. play appeal factors of a game, such as how immersive it is), denoted by subscript, as in \(a_g\) to denote the play appeal factors of a game (forms of \(a\) with different subscript will be discussed later). It can be multi-dimensional to include multiple play appeal factors that are deemed relevant for a game (e.g. immersion, challenge, social bonding, exploration, etc), and is therefore best considered a vector containing play appeal factors, as in e.g. \(a_g = (a_{g1}, a_{g2}, \ldots, a_{gn})\).

**Definition 6:** **Engagement**, denoted by \(e\), is a normalised variable describing the how much the game engages its players (i.e. the amount of cognitive effort and attention directed into the game over time), where zero indicates no time engaged at all, and one is defined as the total amount of engagement needed to fully master the game (and therefore when \(e = 1\), then \(\rho_g = 1\)).

**Definition 7:** A **play theory**, denoted by \(f()\), is a function that uses play appeal factors, \(a\), to make a prediction of engagement, \(e\), and therefore also specifies what play appeal factors are relevant for making that prediction. A play theory specifies the relevant variables to be measured as elements of \(a\), and provides the function, \(f()\), to take those measurements and generate a prediction for \(e\) for an audience:

\[
e = f(a_g)
\]
Where, for a given value of $a$, the function $f()$ must only produce one value for $e$.

**Definition 8:** Learning difficulty of a game, denoted by $c$, is a measure of how difficult the game is to learn. As with $a$, exactly how this variable is conceptualised depends on the theory of learning employed in a particular study, but one simple and useful example conceptualisation of $c$ is as the complexity of the game, $K(m_g)$. Other conceptualisations are discussed in the next section.

**Definition 10:** A learning theory, denoted by $l()$, is a function that produces a prediction of performance in the game, $\rho_g$, based on engagement, $e$, and learning difficulty of the game, $c$. If it does not incorporate individual learning aptitude, and only provides a broad prediction of performance for an unspecified population, it can be written:

$$\rho_g = l(e, c)$$ (3.6)

Where, for any given values of $e$ and $c$, the function $l()$ must only produce one value for $\rho_g$.

Formula (3.6) can be substituted into formula (3.2) as in,

$$\rho_r = l(e, c)\ V$$ (3.7)

Formula (3.5) can be substituted into formula (3.7) as in,

$$\rho_r = l(f(a_g), c)\ V$$ (3.8)

From this it would be theoretically possible to predict the final outcome of real-world learning using the starting measurements of the play appeal factors, learning difficulty, and veracity of the game. But much work remains - namely in
developing the best ways to measure those variables, and developing the theories of play and learning to transform those variables.

Note that MMT defines variables in terms of their relevance to outcomes, not in terms of how they are measured. It specifies the roles that measurable game properties can play in determining outcomes of engagement and learning, but does not demarcate specific game properties into those roles. For example, the game’s informational complexity, $K(m_g)$, could serve in calculating $V$, as in formula (3.1), and/or could serve as a measure of the learning difficulty of the game, $c$, based on a theory of learning that specified its relevance, or $K(m_g)$ could be a play appeal factor, $a_g$, if a study used a theory of play that hypothesised that the game’s informational complexity would affect engagement.

This section provided a broad overview of the abstract mathematics of MMT. The ambiguities of reality that were largely set aside for the above mathematical discussion (e.g. how does one best conceptualise or measure veracity?) will now be addressed in the following section.

### 3.3 Usage of MMT: Measurements & Predictions

#### 3.3.1 Modelling

In MMT, the first process of “Modelling” refers to encoding the real-world domain into a game. This kind of modelling captures what learners need to learn about the real-world domain, but also to produce an engaging game. Therefore, it is not a strictly descriptive modelling process.

It is thus important to distinguish two types of modelling: Descriptive and prescriptive. Descriptive modelling aims to accurately, objectively map its subject matter. For example, a scientist may try to descriptively model a weather system, or a game critic may try to descriptively model a recently-released commercial game - the game critic, just like the scientist, is using descriptive modelling to try to ac-
accurately, objectively describe and thereby understand an existing system that they have encountered.

This is in contrast to prescriptive modelling, which aims to convert the real-world domain into a suitable game. In other words, prescriptive modelling is a specific subset of the game design field: game design when adapting a real-world domain into a game (analogous to adapting a novel to film). For example, a team of educators and game developers may work to design an engaging game of urban development, making specific deliberate deviations from reality (e.g. taking certain liberties or making simplifications to the systems) in a conscious effort to craft an engaging game. A prescriptive modelling procedure instructs such game developers on how to sculpt an engaging game from source material, such as a real-world phenomenon or an educational curriculum. Where descriptive modelling aims to produce an objective, accurate model of a system, prescriptive modelling aims to produce an engaging game based on a system - A distinction similar to a photograph versus a painting. Such prescriptive modelling is the kind referred to by the first stage of MMT.

Rigorous modelling will involve different procedures and guidelines depending on whether it is descriptive or prescriptive. Descriptive modelling procedures should capture as much relevant and accurate information as possible about the subject. In contrast, prescriptive modelling procedures would purport to tell us which elements to omit or alter to convert a domain into a successful (engaging) game. Therefore, many of the frameworks reviewed above that seek to guide educational game design are examples of prescriptive modelling procedures and fall into the first stage of MMT.

Descriptive modelling procedures have been suggested by Sterman (2000), but every model is incomplete as it is an abstract representation of the real world and there is no accepted, correct descriptive modelling procedure to apply to any subject matter. Choosing the level of abstraction at which to model can be a difficult issue (Chu, Strand, and Fjelland, 2003; Grim et al., 2013; Ryan, 2007; Sterman, 2000).
To that end, descriptive modelling procedures will have to be selected or developed to suit a specific purpose.

Many modelling notations have been proposed for games, including Dynamic Causal Nets (Tornqvist, Wen, and Tichon, 2017a), Machinations diagrams (Adams and Dormans, 2012), game atoms (Cook, 2007), and Westera’s (2017) game activity trees. Despite this, there is also no standardised objective formal notation or description procedure for games that could open them up to comparison and mathematical analysis (Araújo and Roque, 2009; Koster, 2005b). The learning theory employed by a particular study can be used to guide the choice or development of descriptive modelling procedure. This will have to be considered alongside a choice of modelling notation (i.e. what format the model will take). For example, a common modelling notation in the cognitive science of play is the Causal Bayes Net (CBN, Gopnik and Schulz, 2007). CBNs (see figure 2.1) are networks of events connected by links representing a causal influence that changes the base-rate probability of the event (for details, see the above literature review, Implicit vs. Explicit Theories of Learning for Games section).

Whatever descriptive modelling procedure is selected for a study, it should capture the elements necessary to determine if learning occurs and transfers, according to the theory of learning employed in the particular study. Ideally the modelling procedures for both the game and the real-world domain would be precise and systematic, such that the resulting model would be reproducible by a different research team working with the same material and the same modelling procedures (for an in-depth discussion on how to make modelling decisions, see Sterman, 2000). Such consistency is necessary to get consistent measures of veracity.

I define veracity as the similarity between the game model and the domain model. When considered in terms of the complexity or informational content, this would be a measure of how much the game simplifies the real-world domain (given that the descriptive modelling procedure should capture what needs to be learned about the real-world domain). The formula for $V$ in the preceding section states that the difference between the game model and the real-world domain model can
PROPOSING THE MODEL-MASTER-TRANSFER FRAMEWORK

provide a measure of the veracity, i.e. the realism or accuracy of the game. If the
veracity is zero, then the game has nothing in common with the real-world domain it
represents. A veracity of 100% would be problematic because that would mean that
there were no differences: It would not be a designed game but merely a descriptive
copy of the real-world domain. Veracity can thus provide some information on how
effective the prescriptive modelling procedure was at producing the intended kind
of game.

Veracity in MMT can be informed by research on fidelity and similar con-
cepts. Hamstra et al. (2014) argue for distinguishing different kinds of fidelity, rather
than using the blanket term due to its vagueness. Physical resemblance refers to
the perceptual similarities, such as look, sound, smell, and feeling. Functional task
alignment is how well the simulation’s functional components demand the appro-
priate skills that are needed to perform the same task in real life. Norman et al. (2012)
used the terms engineering fidelity and psychological fidelity respectively, to refer
to very similar concepts. Both groups of authors stressed the potential ambiguity in
the term fidelity, noting that, “Proximity to ‘real life’ may not be unidimensional”
(2012, p. 637). Therefore, it can be useful to distinguish several dimensions or types
of veracity, depending on the context. Specifically, different learning theories im-
ply different conceptions of veracity. Hamstra et al. (2014, p. 389) notes that, “A
simulator that is considered low fidelity in one circumstance might be considered
high fidelity in another for legitimate reasons”. Therefore, how a study defines and
measures veracity must be informed by the learning theory used to analyse edu-
cational games in each individual study. Conceptions of veracity and learning are
inter-dependent.

Comparing the learning theories reviewed earlier with various literature on
veracity suggests some dimensions of veracity (additional dimensions may come to
light when considering other learning theories, but the diversity of learning theories
is prohibitively vast to cover here):

1. Veracity as Physical Resemblance: Hamstra et al. (2014) define physical resem-
blance as the perceptual fidelity of the artefact, having the same look, sound,
and feel of the real thing. This would be applicable as a form of veracity when using Instance-Based Learning Theory. IBLT posits that learners associate a specific perceptual cue (e.g. falling downward) with a specific motor response (e.g. pulling up on the joystick), building up an instance library of how to respond. This would make physical resemblance critical to effective learning. Therefore, a study that used IBLT to study educational games would best measure veracity in terms of physical resemblance (e.g. the realism of the graphics and control systems).

2. Veracity as Causal Isomorphism: In mathematics, isomorphism is the extent to which two models are structurally identical or mathematically equivalent. If one accepts veracity as entailing isomorphism, one could model one’s real-world domain as a formal system and compare this to the game system. Any features that they have in common bring them closer to being isomorphic. Any differences between the real-world model and the game model would indicate they were not isomorphic, and therefore decrease the veracity score. This is an apt conception of veracity for a causal learning theory such as Causal Bayes Nets. If the game and real-world domain are modelled as networks (as in CBNs), then determining veracity is a matter of counting how many nodes and links are missing or different when the game model is compared to the domain model. More complex measures of isomorphism are also being developed (Schieber et al., 2017).

The concepts of veracity and isomorphism can be explained with an example. Figure 3.2 shows a simplified causal model of an aquaponics system to demonstrate a calculation of veracity as isomorphism. The real-world model on the left is assumed to be complete and correct. It includes seven nodes and eight links. The game model on the right is compared to the real-world model and scored for accuracy. It has seven links that are correct, but one link from fish to carbon dioxide is missing (scoring a zero), and one link from the bacteria to carbon dioxide that is incorrect (scoring a negative one). The seven correct plus the zero for the missing link, and the negative one for the incorrect link, results in a score of six. To get veracity this score is divided by the maximum possible score of eight based on the real-world
model. Therefore, veracity is six divided by eight, which is 0.75. This simple example demonstrates how a veracity score based on isomorphism could be calculated for a game teaching aquaponics based on a causal learning theory like CBNs. Such a procedure would not only be appropriate for teaching aquaponics, but for any studies using simulation games to teach about the structure of systems such as ecology, business, or circuitry (Greiff and Fischer, 2013; Greiff et al., 2015; Ibrahim et al., 2012).

![Figure 3.2: Simplified causal model of an aquaponics system to demonstrate calculating veracity as causal isomorphism. The game model on the right is compared to the real-world model on the left.](image)

\[ 0 \leq V = \frac{1 + 1 + 1 + 1 + 1 + 1 + 1 + 0 - 1}{1 + 1 + 1 + 1 + 1 + 1 + 1 + 1} = \frac{6}{8} = 0.75 \leq 2 \]  

(3.9)

MMT predicts that a game with low veracity is likely to result in lower transfer of learning, due to that difference between the game and the reality. The mastery process mediates this relationship between veracity and transfer.
3.3.2 Mastery

A study investigating mastery should be able to test how much mastery participants achieved after playing the game. This mastery measurement can be normalised, thereby putting it on a universal scale for easy comparison to and integration with other studies. When the player demonstrates a degree of mastery of the game, this could be compared to the perfect and complete knowledge represented by the game, and therefore give a measure of player mastery on a scale of zero to 100%. Exactly how this is measured will depend on the kind of learning in the study. For example, if the player demonstrates full declarative knowledge of all the content presented in the game, they could be considered to have achieved full mastery. Similarly, if the player is consistently executing the optimal strategy flawlessly, then they have mastered the game. In either case, it can be quantified and normalised. Let \( \rho_g \) be the normalised performance of a participant in the game, \( \hat{\rho}_g \) be the raw measured performance of that participant, and \( o_g \) be the optimal performance score (e.g. as measured when an artificial intelligence plays the game optimally):

\[
\rho_g = \frac{\hat{\rho}_g}{o_g}
\]  

(3.10)

Often, performance will be a vector or matrix, containing multiple elements of performance (e.g. subsections of geometry and algebra on a maths exam). Just like the subsections in an exam, each element can be normalised individually, and/or a total score for the entire exam can be calculated. This measure of mastery will be useful for predicting transfer.

3.3.3 Transfer

Much like veracity and fidelity, transfer has proven a difficult concept to precisely define. Barnett and Ceci (2002) characterise transfer research as difficult to interpret, with various successes and failures of transfer in different studies, and attributed this to inconsistency in conceptualisations of transfer. One study might define transfer according to certain factors (e.g. the different environment of the lab vs. the real
world), while another used a different definition (e.g. changing from a written test to a physical performance). Barnett and Ceci therefore break transfer into factors related to content (what is transferred) and context (where it is transferred).

In terms of the context aspect of transfer, transfer is often conceptualised as the distance separating the training scenario from the real-world scenario. This is the distinction between near and far transfer (Barnett and Ceci, 2002; Kiili, 2005; Powers and Brooks, 2014). This distance may be due to any number of differences between the training context and the real-world context, such as modality (written test vs. problem-solving task), or the knowledge domain (carpentry vs. sailing). Note that this concept of the distance between the game and the real-world domain corresponds to MMT’s concept of veracity. Therefore, I will consider a different conception of transfer for MMT.

In terms of the content aspect of transfer, the theory of identical elements states that, “some tasks involve identical processing components. The more identical processing elements two tasks share, the more learning on one will benefit the other — for example, large transfer from learning to drive a car to learning to drive a truck, but less transfer to learning to drive a boat. . . . seemingly different tasks may nonetheless have similar rules at their roots” (Green and Bavelier, 2012, p. 199). Note that the theory of identical elements corresponds almost directly to veracity conceptualised as isomorphic rules, suggesting that physical resemblance will not encourage transfer, as supported by some studies on fidelity (Hamstra et al., 2014; Norman, Dore, and Grierson, 2012). Thus, theories of learning, transfer and veracity are often deeply interdependent.

Transfer is much like veracity in MMT: There is an ongoing debate as to how best to define it, and MMT does not aim to resolve that debate. MMT aims to be agnostic towards what theories are best applied to the sub-processes of modelling, mastery and transfer. However, it is clear from the work in these areas that theories of learning, transfer and veracity will be deeply inter-dependent, and therefore selection of one should inform the others.
The above quotation from Green and Bavelier focuses on the content aspect of transfer. They use the term ‘transfer’ to refer to the magnitude of the learning that is retained between the two scenarios (e.g. the training, and the real world), rather than the distance between the scenarios. Because the context aspect (distance between the scenarios) is already captured by veracity in MMT, the below equations will attempt to capture the content aspect (how much learning is retained between the two contexts).

Once measured, the amount of mastery can be multiplied by the veracity of the game to give a prediction of the amount of transfer, and therefore how well they will perform in the real world. For example, when the game’s veracity is 100%, and the learning is 100%, one would expect performance in the real world to also be very high. If either of those are zero (e.g. the game has nothing in common with the real-world domain, or the player has learned nothing), then predicted transfer is zero. In reality, both variables are likely to be somewhere in between, predicting a particular degree of transfer. Note that transfer entails performance, and therefore could be a vector containing all the same elements of the game performance vector. Let $\tilde{\rho}_r$ be normalised predicted real-world performance for a participant, $\rho_g$ be normalised measured game performance of a participant, $V$ be game veracity:

$$\tilde{\rho}_r = \rho_g V$$  \hspace{1cm} (3.11)

Calculating this predicted level of performance allows a comparison to the actual measured performance as a test of the reliability of MMT. In order to make such a comparison, real-world performance would have to be normalised as well, which would be done in much the same way as game performance. Let $\rho_r$ be the normalised performance of participants in the sample, $\hat{\rho}_r$ be the raw measured performance of participants, and $o_r$ be the optimal performance score:

$$\rho_r = \frac{\hat{\rho}_r}{o_r}$$  \hspace{1cm} (3.12)
PROPOSING THE MODEL-MASTER-TRANSFER FRAMEWORK

The measured real-world performance can then be compared to game performance to see how much of their mastery of the game carried over to the real world, thereby giving a measure of transfer. Let \( T \) be transfer for the sample, \( \rho_r \) be the normalised measured real-world performance of the sample, and \( \rho_g \) be the normalised measured game performance of the sample:

\[
T = \frac{\rho_r}{\rho_g}
\]  

(3.13)

Note that this formula is just an algebraic re-arrangement of formula (3.11) for predicting real-world performance based on game performance and veracity. Therefore, MMT entails the following axiom:

\[
T = V
\]  

(3.14)

This could be a crucial component in testing the validity of a game, or MMT itself, or just a way of triangulating variables from multiple approaches.

Now that I have explained the broad mathematical relationship between modelling, mastery, and transfer in MMT, I will elaborate the “play” and “learn” aspects of mastery, each in turn. This will detail the mathematical relationship between the properties of the game and predictions of mastery.

3.3.4 Engagement & Play Appeal

Engagement being a core concept in game design, I now formally describe its relationship to the other variables that have been established above, thereby demarcating engagement’s place in MMT. Here, engagement will be used in the literal sense of engaging with an item: attending to and physically or cognitively manipulating it. Note that this is a behavioural definition of engagement concerned with what one actually does. For example, it is possible to be engaged but experiencing negative affect. Although positive affect and motivation are likely to often be correlated with engagement.
Time spent playing is a direct measure of engagement. It is often treated as an indicator of liking a game (Johnson et al., 2014; Malone, 1980). For example, in many studies on intrinsic motivation, free time spent on an activity is often used as a measure of intrinsic motivation (Boggiano and Ruble, 1982; Cameron, Banko, and Pierce, 2001; Elliot and Harackiewicz, 1996; Lieberoth, 2015; Ryan and Deci, 2000). Free time spent could also be useful in evaluating the educational efficiency of a game by comparing free time engaged to the time required to achieve full mastery.

Engagement has a dimension of intensity and one of time. A game might engage players with a mild degree of intensity, but maintain engagement consistently over a long period. Another game might engage players with a fiercely high intensity that only lasts briefly. Therefore, engagement is more fully understood when it is considered as magnitude over time. This becomes very relevant for educational games where predicted learning could be a function of engagement over time and the difficulty of the content to be learned. The more difficult the content, the slower the learning, and therefore the more time they must spend engaged.

A total amount of engagement that has occurred can be calculated as the area under the engagement-over-time curve. An educational game designed to deliver a lesson once is more likely to be interested in total engagement. But there are often valid reasons to distinguish between intensity and duration. A serious game to encourage exercise and a healthy lifestyle is going to be much more interested in maintaining engagement over a long period.

Predicting engagement depends on theories of play motivation (e.g. challenge, immersion, etc.). Whatever theory of play one adopts, one would predict that certain game designs will result in more or less engagement due to how successfully they appeal to that theorised play motivation. For example, working from immersion, one would predict that a game that creates a deep sense of immersion would be more engaging than a game with weak or broken immersion, and “immersion” could be operationally defined based on a specific theory thereof.

Crucially, such definitions can and should be developed into quantitative operational definitions so that researchers can quantify the strength of this play ap-
peal in different game designs and produce more precise predictions. For example, there have been proposed methods for quantifying the degree of challenge in a game (Aponte, Levieux, and Natkin, 2011; Fraser, Katchabaw, and Mercer, 2014; McMillan, 2013; Tornqvist and Tichon, 2019). These precise quantitative measurements of the play appeal of a game would allow a study on the first process of MMT – modelling the domain into a game – to apply its game design framework (a prescriptive modelling procedure) to produce a game, then objectively quantify the game properties relevant to a theory of play, and therefore test how successful their prescriptive modelling procedure was at producing a game with the intended properties.

Precise objective measures of play appeal factors are also vital for populating these formulae and testing theories of play. The amount of play appeal in a game should be quantified so that it can be used to provide predictions of engagement, because engagement is relevant for predicting mastery.

Play appeal can be a vector, as (dependent on the theory of play) there may be multiple parameters that interact to give rise to play appeal. Let \( \hat{e}_p \) be predicted engagement of players, \( a_g \) be a vector containing the play appeal parameter/s of the game, and \( f(a_g) \) be the function described by the theory of play that explains how the particular game parameter/s relate to engagement (some theories may purport a linear relationship, but others may describe a more complex function):

\[
\hat{e}_p = f(a_g)
\] (3.15)

This would be used for research into general effects across populations. But studies on individual differences would want to include parameters of an individual person (i.e. personality) to come up with an individualised prediction for that person. Let \( \hat{e}_p \) be the predicted engagement for a specific person, \( a_p \) be a vector containing the play appeal parameters relevant to the specified person (the individual differences in play preferences for this particular individual), and \( f(a_g, a_p) \) be the function describing how the theory of play relates those parameters to engagement:
\[ \hat{e}_p = f(a_g, a_p) \] (3.16)

For example, where \( a_g \) may include a measure of challenge in the game (e.g. the number of enemies present), \( a_p \) would include a measure of how much a specific participant prefers to play games for the sake of challenge (e.g. measured with a questionnaire), and therefore how much challenge is likely to engage that particular participant (the higher both play appeal factors, the higher the resultant engagement). Or, \( a_p \) could include a measure of the participant’s preferred level of challenge, in which case the function, \( f() \), to predict engagement may entail a comparison of the two values to see how well they matched, rather than just multiplying them together.

These predicted levels of engagement could be compared to the actual, measured levels of engagement to test the accuracy of the theory of play.

If pilot tests are able to determine how much time engaged is needed to fully master the game under controlled conditions, then measured engagement can be normalised. Let \( e_p \) be normalised engagement for a participant, \( \hat{e}_p \) be engagement measured for a participant during the actual experiment, and \( e_g \) be the engagement needed to achieve complete mastery according to pilot studies:

\[ e_p = \frac{\hat{e}_p}{e_g} \] (3.17)

This measure of engagement can now be used to predict how much of the game players will master. This requires a measure of how much there is to learn, or how difficult it is to master.

### 3.3.5 Learning Difficulty & Complexity

A game that has more to learn, or is more difficult to learn, will require more time engaged to fully master. One way to conceptualise learning difficulty is as a form
of complexity. The complexity of the game (quantified with e.g. Kolmogorov complexity) can be used to give a prediction of the amount of game mastery to expect. It would have a direct effect on the difficulty of learning, and therefore time required for mastery would increase with complexity. More complex should mean more difficult to learn. But it might also have an indirect effect on learning through play appeal and engagement. For example, according to Flow theory (Czikszentmihalyi, 1990), player engagement will be highest when the game is not too simple as to be boring, and not too complex as to be frustrating. Therefore, complexity could be one of the play appeal factors included in \( a_g \) for predicting engagement based on a theory of play in equations (3.16) and (3.17). Like veracity, how complexity is measured should be derived from a theory of human cognition.

Predicted game mastery is a function of engagement over time (the longer one spends on something, the more they are likely to learn), and difficulty of learning (complexity, in this example). Let \( \tilde{\rho}_p \) be predicted normalised average mastery for the sample, \( c \) be game complexity (or amount of content to be learned, or any other measure of learning difficulty), and \( e_p \) be normalised measured average engagement of the sample (one could also substitute in \( \tilde{e}_p \) as the predicted normalised amount of engagement if one lacked access to an actual measured value), and \( l(e_p, c) \) be the function relating these variables to mastery as described by the theory of learning used in the study:

\[
\tilde{\rho}_p = l(e_p, c) \tag{3.18}
\]

Studies investigating individual differences would want to include parameters of individual learning aptitude to generate an individualised prediction. For this, one would need some relevant information about the cognitive aptitudes of the individual learner to inform a prediction of how much they might struggle or excel with the learning material. Such cognitive aptitude variables could be as broad as a measure of their general intelligence or their general curiosity, or it could take the form of more specific measures of their background knowledge and interest in related subject matter (e.g. a marine biology enthusiast may do much better than a
PROPOSING THE MODEL-MASTER-TRANSFER FRAMEWORK

novice in mastering an aquarium game). Let $\tilde{\rho}_p$ be predicted normalised mastery for a participant, $c$ be game complexity (or other measure of learning difficulty), $e_p$ be normalised measured engagement of a participant, $\zeta$ be a vector containing factors of cognitive aptitude for the individual participant, and $l(e_p, \zeta, c)$ be the function relating these variables to mastery as described by the theory of learning used in the study:

$$\tilde{\rho}_p = l(e_p, \zeta, c)$$  \hspace{1cm} (3.19)

These formulas show how MMT could be used to break educational video games into more specific sub-processes to study in greater depth and precision individually. They also specify how such narrow studies could be integrated with others to build up a larger causal picture of how educational games achieve positive outcomes. A game design technique (i.e. a prescriptive modelling procedure) may produce games with certain properties (play appeal factors, learning difficulty, and veracity), and certain game properties result in some degree of engagement, which results in some amount of game mastery, which yields a certain amount of transfer to the real-world domain, depending on the veracity of that game. The overview of figure 3.1 is expanded with these added details as a summary for quick future reference in figure 3.3.

Figure 3.3: Summary of the variables and formulae proposed here.
MMT has interesting implications and applications for not only researchers, but also practitioners such as game developers and educators.

3.4 Implications & Applications

MMT is a tool for developing a better scientific understanding of how educational video games yield positive outcomes. But there are also useful ways it can be applied by practitioners, such as educators and software developers who are using educational video games.

3.4.1 Implications for Practitioners

For Educational Game Developers:

1. Select a specific prescriptive modelling procedure (i.e. a game design framework), reporting what procedure was used and what the outcome was (see points below for reporting outcome). This will help establish what kinds of prescriptive modelling procedures produce what kinds of games.

2. Report how the design of specific game properties were derived from a specific theory of play/learning. Or, if designing informally through iterative playtesting, then still measure and report game properties (see next points).

3. Quantify Game Properties: In the resulting game, objectively quantify properties that are purported to have a causal relationship to engagement \( a_g \) or learning \( c \) according to your selected theory of learning/play \( f(a_g) \) or \( l(e_p, c) \). This will help determine the effectiveness of the prescriptive modelling procedure. For example, a prescriptive modelling procedure might be built to produce the ideal amount of challenge, or to maximise veracity. Therefore, those game properties need to be quantified and reported to have a record of the success of the procedure.

   (a) Veracity: Report the veracity of the resulting game, e.g. by using a descriptive modelling procedure to model both the game and the real-world
domain on which it is based, then applying a measure of veracity, fidelity, and/or isomorphism, derived from your selected learning theory.

4. Adjustable Game Properties: Academics can test theories of play and learning, discovering the relationships between specific game properties and outcomes such as learning and engagement. But they can only do so if developers build into their games administrative controls to adjust the levels of different game properties (e.g. $a_g$, $c$, or $V$). For example, there could be a configuration file that scientists or educators could edit to manipulate the degree of challenge, complexity, or other properties in the game to test different theories of play. At its most basic level, a feature such as multiplayer functionality could be simply turned on or off. Deciding which game properties to open up to experimental manipulation should be based on the theory of play/learning used to design the game (e.g. if the game design was primarily concerned with maximising player agency\(^1\), then agency should be an adjustable parameter). This feature is not just useful for scientific investigations, but also allows the game to be adapted to different contexts as new research comes to light concerning what combinations of what levels of which game properties maximise engagement and learning.

5. Build in Measures of Mastery and Engagement: Build into the game systems that measure the engagement and mastery of players. These can either be explicit questionnaires, or subtle, invisible measures built into the game code itself that simply record natural player behaviour to derive measures of engagement (e.g. time spent playing) or mastery (e.g. how optimally they are playing, as in eq. 3.12, or how high they are scoring).

6. Report Mastery and Engagement Outcomes: Based on some form of playtesting or actual field use, record and report the levels of engagement and mastery for players. This helps test the success of the prescriptive modelling procedure, and can be used to derive predictions of transfer using the discussed equations (e.g. eq. 3.18 and 3.11).

\(^1\)Quantifying agency is discussed in the chapter *Study 1: Evaluating Play Appeal Factors & Engagement*. 

94
PROPOSING THE MODEL-MASTER-TRANSFER FRAMEWORK

For Educators:

1. When working with game developers to produce a game, employ the approaches listed above for game developers.

2. Instead of (or in addition to) comparing the game intervention to a no-game intervention, compare the game to itself with its game properties adjusted for different groups (see point 4 for game developers, above). For example, compare a group using a low-challenge version of the game to a group using a high-challenge version. Measure and report which levels of the different properties (e.g. $a_g$, $c$, or $V$) were used for different groups.

3. Measure Engagement and Mastery: Measure and report students’ engagement with and mastery of the game. This is a direct test of the effectiveness of the design of the game, but is also necessary to test theories of transfer - We need to know what percentage of their game mastery is carried over to the final test (e.g. the academic exam, or the real-world problem-solving task), as seen in eq. 3.13.

4. Report Transfer: When you test for improved academic outcomes or performance ($\rho_r$), compare this to the amount of game mastery ($\rho_p$) to determine how much mastery was able to transfer to the real world (eq. 3.2). For example, if improved real-world performance is uncorrelated with improved game mastery, then it may simply have been the enjoyable experience of the game, and not its educational subject matter, that benefited students.

5. When trying to maximise educational efficiency, you can use MMT to catch early any interventions that are not likely to succeed, to spend more time on interventions more likely to succeed. For example:

(a) When using a theory of play/learning to create a game with particular play appeal factors, MMT suggests you should quantify the properties of the game that correspond to those play appeal factors (such as Kolmogorov complexity, or degree of challenge). If the game design process failed to produce a game with the intended properties (e.g. $a_g$ is much
lower than the intended $a_d$), then this can be caught early and you can try again with a different prescriptive modelling procedure, rather than going through the whole process of testing the game and finding it did not incite engagement.

(b) Before you use a game in the classroom, if you can get a measure of its veracity (derived from a theory of learning), then you can get an upper limit on the usefulness of the game (eq. 3.11): If veracity is high, then it is possible to have large transfer benefits. If veracity is low or zero, then (according MMT’s equations) that means anything learned in the game is unlikely to have any bearing on the real-world domain you are trying to teach. Therefore, a different game should be used.

(c) When students are playing the game, you should measure their degree of mastery of the game ($\rho_p$). If some or all students are only achieving low levels of mastery of the game, then (according to the eq. 3.11) there is likely to be little or no improvement when it comes to transfer that mastery to the real-world domain. Therefore, the intervention can be ended right there and you can try again with a different approach, without necessarily having to actually do the transfer tests. Catching failed interventions early like this can save time and thereby increase the total number of different interventions that can be trialled.

### 3.4.2 Implications for Researchers

MMT contributes a framework that facilitates deconstruction of educational video game research into more specific sub-processes, and facilitates integrating studies on such sub-processes, assembling them to gradually build up a larger causal model of how educational video games achieve positive outcomes for players. This has important implications for the practice of researchers in this field:

1. It would be invaluable to narrow the focus of your study to investigate just one or two processes at a time with greater depth and specificity. For example, just investigating transfer (eq. 3.11), or just investigating mastery (eq. 3.18).
2. Narrow the focus to the relationship between several specific variables, deriving hypotheses from specific theories of play or learning. For example, testing the theory that increasing complexity increases challenge and therefore engagement, but decreases mastery (the hypothesis that a single game property - complexity - can operate as $c$ and $a_g$ in eqs. 3.16 and 3.18).

3. Use or develop games that allow you to manipulate these variables.

4. Use or develop objective measures for these variables, based on the theories of play or learning used in the study (e.g. quantifying agency in a game as the number of options available to the player at a time).

5. Where possible, normalise variables to put them on an absolute scale (see eqs. 3.1, 3.9, 3.10, 3.12, 3.17) to enable inter-study comparison and integration into MMT to build up a larger causal model. For example, compare the time participants spend playing with the amount of play time required to master the game, thereby normalising the engagement variable.

6. If investigating the first process - modelling the domain as a game - then select or develop a specific prescriptive modelling procedure (a game design framework), follow it, and report on the outcome in terms of the properties of the resulting game, and in terms of the repeatability of the procedure (e.g. does a different team following the same procedure produce a similar game?).

7. If investigating the final process - transfer - then use or develop a theory of learning and transfer to derive hypotheses of how game veracity and players’ mastery of the game will affect real-world performance (eqs. 3.18 and 3.11).

MMT’s implications for research are best explained with an example. Here I will consider previous research developing tools to study challenge and flow in games, and how MMT delineates their relevance to other studies on similar or distinct topics.

Previous literature reviews and meta analyses (e.g. Boyle et al., 2016; Clark, Tanner-smith, and Killingsworth, 2014; Connolly et al., 2012) have recommended
research shift from broad proof-of-concept studies to narrower studies of how specific educational video game properties affect specific outcomes of engagement and learning. There are already many tools that have been proposed to study these individual pieces of the causal picture, such as Adams and Dormans’ (2012) game modelling notation, or Aponte et al.’s (2011) method for quantifying challenge. But most such formal tools have not yet seen widespread use in the education or educational video games field. However, such tools are invaluable for studying narrower sub-processes.

In the above literature review, I detailed how there have been many frameworks for educational games proposed, but I could not locate one that sought to facilitate the integration of such narrower research into a cohesive whole. MMT allows a more detailed causal picture to be incrementally built up from multiple narrow studies that each investigate a smaller piece of the picture.

MMT specifies how studies proposing formal tools such Adams and Dormans (2012) and Aponte et al. (2011) can be employed to contribute to understanding how educational video games produce positive outcomes. When viewed in isolation, each of these studies might be easily dismissed as irrelevant to designing educational video games. But how they fit within MMT to contribute to this field is clearer when running through an example of how they can be strung together to span the entire, larger picture:

1. Modelling: To investigate the theory of flow, a prescriptive modelling procedure (e.g. that of Zee, Holkenborg, and Robinson, 2012) can be used to design a game to teach about, for example, business management, using a formal notation system (e.g. that of Hunecker, 2013) to document and report the structure of the game. That formal description of the structure of the game can then be compared to an expert’s formal diagram of the real business made using the same formal notation system, to derive a measure of isomorphism to report a dimension of veracity for use in MMT’s equations (see e.g. eqs. 3.1 and 3.9). Then a method to quantify the degree of challenge in the game (e.g., that of Aponte, Levieux, and Natkin, 2011) can be used to determine if the prescrip-
tive modelling procedure produced the intended degree of challenge \((a_g)\) to engage players (see section *Engagement & Play Appeal*).

2. Mastery: Then, a tool (e.g., Sweetser, Johnson, and Wyeth, 2012; Sweetser and Wyeth, 2005) can be used to measure outcomes of flow to confirm if the intended degree of challenge causes the intended experience of flow, corroborated with other measures of engagement (e.g., Rigby and Ryan, 2007; Ryan, Rigby, and Przybylski, 2006) and free time spent playing (eq. 3.15). Then players’ mastery of the game can be quantified as their performance relative to the optimal possible performance in the game (eq. 3.10). That measure of mastery can be used with the measure of veracity to generate predictions of transfer (eq. 3.11).

3. Transfer: Then players can be tested on a real-world transfer task. This not only provides an overall measure of the success of the educational video game, but can also feed back into the definition and theory of transfer and veracity – If these equations predicted a different amount of transfer to what actually occurred (eq. 3.11), this brings into question the definition or measures used to quantify veracity, and/or its relationship to mastery and subsequent transfer.

This above example demonstrates how a very narrow study dedicated to establishing a mathematical way to measure challenge, such as Aponte et al. (2011), even though it could be easily dismissed as having limited direct use for someone who is, for example, studying transfer, it actually addresses a key step in the causal chain that ultimately determines the success of an educational video game.

It would be unrealistic for a singular study to attempt all of the above steps across modelling, mastery, and transfer. But the value of MMT is to facilitate narrower studies that each examine one or two of the above parts of the process, while maintaining relevance to the larger picture. For example, a study that just used a prescriptive modelling procedure to design the game, then used a method for measuring its degree of challenge, provides vital information for designing educational video games when it is inserted into MMT. MMT demarcates the exact role and relevance of such narrower studies to the field at large by specifying exactly how
they can be assembled to form a coherent causal chain toward the ultimate outcome of transfer to the real world. Thus, narrower and deeper studies are able to build directly on top of each other’s previous work, and disparate studies on distinct processes are able to inform each other to contribute to a broader and deeper understanding of how these processes interact.

3.5 Limitations & Future Work

MMT is limited to a focus on serious games, and particularly educational games, but is modular enough to be easily elaborated and extended to improve its relevance to, for example, non-serious entertainment games. Specifically, any elements of MMT concerning engagement (as in studies one and two of this thesis) will naturally have relevance to non-serious entertainment games, although the model would have to be extended to specify how engagement relates to game sales or word-of-mouth marketing.

This proposal of MMT is limited in that it is only theoretical. The specific predictions made possible with the above mathematical expression of MMT could be tested experimentally to determine if these equations bear accurate predictions or require reworking. Furthermore, MMT (by design) only describes the entities and processes at a high, abstract level. It therefore depends on additional theoretical tools to fill in the details – for example, a theory of play to specify the exact relationship between play appeal factors and engagement, or a descriptive modelling procedure to help ascertain a measure of veracity. Consequently, MMT inherits the limitations of such tools. For example, veracity and transfer are much-debated concepts in need of further refinement (as explained above). Similarly, the current lack of any standardised notation for game modelling (Araújo and Roque, 2009; Koster, 2005b) or replicable descriptive modelling procedure also limits the usefulness of MMT, and therefore is a crucial area for future research.
3.5.1 Modelling

Prescriptive modelling procedures (e.g. Zee, Holkenborg, and Robinson, 2012) need to be developed and tested for several outcomes: Their consistency to produce similar game designs given different design teams working with the same source material and prescriptive modelling procedure to design a game; The degree of veracity the procedure tends to achieve; The properties of the games the procedure tends to produce - for example, a prescriptive modelling procedure might be designed specifically to achieve the optimal amount of challenge in the resulting game design. Therefore, the level of challenge in the resulting game should be measured and quantified to test if the procedure was successful in its aims. This knowledge can then be integrated with studies testing different levels of challenge to determine which result in the most engagement for players.

3.5.2 Mastery Through Play

To that end, much work is needed to formalise the myriad theories of play motivation such as challenge. These play appeal factors need to be developed into more precise models that specify exactly what quantifiable properties of games are relevant to outcomes such as engagement and learning, and thereby generate specific predictions of these outcomes using measurements of play appeal factors in specific games. This requires not only theoretical work on the nature of these play appeal factors, but also empirical work to test their predictions and tweak the models to produce accurate and reliable predictions. These can then feed back into work on prescriptive modelling procedures by indicating which game properties need to be carefully designed and in what way. Such work will also feed back into theories of individual player preferences, playstyles, and learning aptitudes, that can provide more individualised detail on how engagement and mastery occur.
3.5.3 Transfer

Finally, theories of transfer and veracity need further empirical and theoretical work, likely alongside descriptive modelling procedures to ensure the source material and the game are modelled in a rigorous and consistent way that should theoretically capture the elements of those systems that are relevant to outcomes of engagement, learning, and transfer. Then, when enough studies have taken the care to measure and report the needed variables, they can be assembled using MMT to gradually build up a quantitative model of how educational and serious games achieve positive outcomes.

3.5.4 Social Context

The least frequent element identified in the literature review of frameworks (see above section, Entities & Processes in Existing Frameworks) was the social context. This is perhaps due to the fact that not every serious game is a multiplayer game, or likely to be used in a social context (for example, it may be played in isolation on public transport), and therefore the social context element only comes into play for a subset of serious games, not all serious games. Nonetheless, ways to elaborate MMT to incorporate the element of social context would be worthwhile. The progress so far on these factors is included in appendix A, Massively Scalable Learning: The benefit of a serious game can be multiplied by increasing its reach to a larger population, and that effect can be decomposed into learnable, practical, and motivational elements operating at both the individual and population levels. This is part of ongoing work to consider the factors at play at the individual and population level to affect how much benefit a serious game can produce, as a product of its benefit on the individual level multiplied by the number of people reached (i.e. it’s saturation) across the target population. This required considering what design features can serve as impediments or incentives to share games. More work is needed to incorporate these variables into MMT.
3.6 Conclusion

Multitudinous studies on educational and serious games have together produced somewhat mixed results (see e.g. Clark, Tanner-smith, and Killingsworth, 2014; Ibrahim et al., 2012; McClarty et al., 2012; Morris et al., 2013; Ronen and Eliahu, 2000; Wouters et al., 2013). Some educational video game projects yield significant measurable benefits for their players, while others do not. The meta analyses and literature reviews gathered in this paper recommend that future work narrow its focus to more systematically investigate how specific design changes (e.g. changing measurable game properties) affect specific outcomes such as engagement and learning (Boyle et al., 2016; Clark, Tanner-smith, and Killingsworth, 2014; Connolly et al., 2012; McClarty et al., 2012). This led to RQ1: How can serious games research move from proof-of-concept studies towards building up a formal model of how specific design changes affect specific outcomes such as learning and engagement?

Unable to locate an existing framework that facilitates such an approach to educational video games research, I proposed my own: MMT. This new framework forms the primary contribution of this research project and helps demonstrate that educational game research can be divided into smaller subprocesses that can be studied in greater depth while still retaining relevance to the larger picture of how to design effective educational games, when that research occurs within a unifying framework.

The Model-Master-Transfer framework conceptualises this field in terms of three entities – the game, the player, and the real-world domain – and their relationships. Many educational video game studies bundle together all three processes of modelling, mastery, and transfer (Barton et al., 2016; Cheng and Su, 2012; Hsu, Tsai, and Wang, 2012; Sengupta, Krinks, and Clark, 2015; Squire et al., 2004), rendering it impossible to determine where things went wrong if the results are negative. But if a study investigates only one process, it can be easily overlooked for having insufficient relevance to the larger field of educational video games. MMT specifies the broader relevance of narrower studies. It provides a structure by which to break down educational video game research into more specific sub-process to study in greater depth, and also detailing how such narrow studies can be assembled to grad-
ually build a larger, more comprehensive causal model of how educational video games achieve positive outcomes for players.

This requires considering the variables that serve causal roles in the success of each of the three processes, and how those variables relate mathematically. MMT proposes such formulae. Successful modelling should result in high game veracity, as calculated by the difference between the real-world domain model and the game model. Mastery is predicted to be determined by the difficulty of learning (e.g. game complexity) and player engagement, and measured in terms of how the player’s performance differs from perfect performance derived from the game model. Transfer is predicted to be a function of the veracity of the game and the achieved level of mastery, and measured in terms of the difference between the player’s performance in the game versus their performance in the real-world domain.

MMT has many practical implications. For researchers, properties of the game that are causal variables in a theory of play or learning should be quantified to derive predictions. For example, if the theory of play hinges on agency, then use an operational definition of agency to quantify the degree of agency present in different experimental conditions to make predictions of engagement. For educational video game designers, they should build into their games measures of engagement and mastery, and ways for educators and researchers to manipulate properties of the game to test hypotheses and adapt the game to future contexts and findings. For example, there could be a configuration file that educators can edit to adjust the level of challenge or agency in the game. For educators, MMT reveals ways to improve the efficiency of trialling multiple game interventions, by checking the game properties like veracity of the game, and the level of mastery of players, to catch early trials that are likely to fail, to save time to try more promising games.

Many of these recommended practices (quantifying game parameters relevant for play appeal, measuring complexity) could be facilitated by a model of the game. Making sense of mixed results would be aided if researchers would not just briefly describe their games, but model them with formal notation and provide the model diagrams as appendices in their studies.
There are still many obstacles and sources of ambiguity in the field of educational video games research. Some could be combatted with the recommendations provided in this chapter. The advancement of educational video games necessitates deeper investigation of the processes of modelling, mastery and transfer, more consistent and more detailed reporting of objectively quantifiable parameters, variables and models of games. Regardless of whether MMT proves popular, hopefully some of these recommended practices will become common.

The next chapter is a contribution detailing one way to achieve the modelling portion of MMT. As previously explained, there is no universal modelling method for all kinds of systems or games and therefore methods will have to be selected and justified based on the context of each study. But the following method provides one option that can be used, and serves as an example of how modelling for MMT can be done to inform the selection or development of other modelling methods, and how to use it to measure a game’s Kolmogorov complexity.
Chapter 4

Proposing Dynamic Causal Nets: Calculating Game Kolmogorov Complexity

4.1 Introduction

The following is an extensive elaboration of a published conference paper, with some passages retained verbatim (Tornqvist, Wen, and Tichon, 2017b, © 2017 IEEE).

This chapter focuses on the modelling component of MMT. The previous chapter proposing MMT provided an overview of all the complexities and concerns within the modelling component of MMT, such as the definition of model veracity, and the diversity of modelling methods in existence, with no established "correct" or universal modelling method. This chapter elaborates the modelling component of MMT by proposing a suitable modelling method for use with MMT, which may be itself used in a study or simply serve as a reference point to help select or develop an appropriate modelling method in future studies. This chapter also demonstrates how a modelling method can be used to measure the complexity of a game, as would be useful in MMT (or any study concerned with game complexity).
Therefore, this chapter contributes methods to objectively quantify several properties of games and their features that are relevant to outcomes of engagement and learning, suitable for use in the MMT framework and others like it. Methods to objectively quantify properties of games will be instrumental in establishing connections between specific game properties and specific outcomes such as engagement and learning. Thus, contributions such as these augment the usefulness of MMT, as well as being contributions in their own right.

I have proposed a modelling notation based on how games are commonly diagrammed in the field: as network of entities ("nouns") connected by interactions ("verbs"), combining conventions of causal Bayes nets and stock and flow diagrams. I termed this notation Dynamic Causal Nets (DCN), and explained how the Kolmogorov complexity of a game could be calculated from its DCN diagram (Tornqvist, Wen, and Tichon, 2017b). Tools such as these empower MMT. For example, as detailed in the above section, Mathematical Structure Overview, the complexity of the game could be used to estimate veracity, and/or serve as a measure of the learning difficulty presented by the game, and/or even function as a play appeal factor affecting engagement.

4.2 Proposing Dynamic Causal Nets

DCNs are based on some of the more common system notations used in the game design and cognitive science literature. As detailed in the literature review, various cognitive studies have used causal Bayes nets (CBNs) as effective models of learning about unfamiliar systems (Gopnik, 2011; Gopnik and Schulz, 2007; Gopnik and Wellman, 2012; Schulz, Standing, and Bonawitz, 2008; Sobel, Tenenbaum, and Gopnik, 2004). CBNs form a network where each node is an event, and each link is an influence on the probability of another event, changing its base rate probability (see figure 2.1 and above literature review for details, Implicit vs. Explicit Theories of Learning for Games section). CBNs form an attractive starting point for game system modelling due to their usage in studies on learning through play, which involve learning causation through interaction with a system (see above citations).
It is perhaps no coincidence that actual game design and analysis is frequently done using similar causal networks to those found in the cognitive science of learning through play (e.g. Adams and Dormans, 2012; Smith and Smith, 2004; Zupke, 2016). Causal networks seem to be an intuitive method for formally describing interactive systems to open them up to mathematical analysis. Many diagrams produced by game designers (e.g. Zupke, 2016) look very similar to CBNs that could have been made by cognitive scientists who might study learning in such games (e.g. Gopnik and Schulz, 2007). One major difference is that most cognitive science CBNs represent changeable states or events as nodes, whereas these example game design networks represent “nouns” as nodes (a “noun” is simply any recognizable game entity, and it is contrasted with a “verb”, which is an interaction between nouns).

For example, it might represent a worker as a node, raw material as another node, and a finished product produced by the worker from the material as a third node, each connected by links representing verbs such as "collect", or "build" (figure 4.1). But there is no standardisation of notation in the industry and thus even these basic conventions are often bent or broken for a specific diagram, such as including an abstract concept or a process as a node. Note that such game design diagrams could not be called CBNs because they represent a different mathematical structure from CBNs and thus are not directly interchangeable, despite having some similarities. For example, a CBN would need to use the events of "build" or "take" as nodes instead of links.

Figure 4.1: Simple example noun-verb causal net as one might find in the game design literature (e.g. Zupke, 2016) shows similarities to CBNs (figure 2.1) but also differences such as often using nouns as nodes and verbs as links.

This shift of focus from states and events to nouns and verbs creates other differences in the modelling methods of traditional CBNs versus game design causal
nets. For example, CBNs are always acyclic, meaning that the links of causation are not permitted to loop back around such that a state is able to (directly or indirectly) affect itself. In CBNs, the flow of cause and effect must always march forward towards the ultimate final state we are interested in, and never create feedback loops. This restriction is problematic for systems that include cycles and feedback loops, which includes many games (Adams and Dormans, 2012), and of course various real-world phenomena. By permitting cycles, the noun-based causal nets of game design become capable of simulating dynamic and chaotic systems. Therefore, to mark this distinction I will refer to these noun-based nets as Dynamic Causal Nets (DCN).

Comparing the subjects of analysis in these different domains reveals another critical reason for the difference. In most of the cognitive science research, the systems are kept relatively small and simple in order to control variables and so that they can be easily learned in the short timeframe of a laboratory experiment. When dealing with such small systems, it is practical and useful to represent state changes as nodes and probabilities as the links between them. In contrast, many of the game design diagrams are of large, complex game systems, involving subsystems that could be broken down even further. When dealing with such large systems, it is more practical to have the nodes refer to nouns, and let the state changes be implied by activity across the network.

Therefore, the dynamic causal networks of game design are able to keep their overall size and complexity more manageable by employing these conventions. Another important convention is that the nodes represent different kinds of nouns, not specific instances of nouns. For example, there might be a node to represent a dog, but not a node to represent a particular dog, Rover, because Rover (being a dog) would have all the same links as every other dog. The links between nodes are the verbs, which describe the interactions between nouns. For example, maybe a dog has a 30% chance of eating a rabbit if it encounters one. This interaction would be represented with a link between those two nouns with a certain weight to convey the strength of that interaction (30%) and its negative effect (destroying the rabbit). In contrast, when a rabbit encounters another rabbit, it has a high chance of breeding
and making a new rabbit. Therefore, the rabbit node would have a link connecting back to itself, with high strength and positive effect (figure 4.2).

Figure 4.2: An example of a simple DCN, showing the current number of each noun (which will change over time as it is simulated), and their interactions as weighted, directed links that can affect the number of that noun in the world, which in turn affects the probability of nouns encountering each other and interacting further. Figure reproduced with permission from Tornqvist, Wen, and Tichon, 2017b, © 2017 IEEE.

Further, the size of a node can be used to represent the density of that noun’s population in the world (e.g. a high population of rabbits in a small space), and thereby represent its likelihood to interact with other nouns. This variable, noun density, is the product of the number of the noun multiplied by the base rate probability, and divided by the size of the game space (i.e. five rabbits in a shed are likely to frequently influence each other, whereas the same five rabbits in a vast open forest are comparatively very unlikely to meet each other very often). This base rate probability can represent physical size, or sphere of influence. For example, if the noun is relatively large like a dog, it might only take 1 or 2 of them to have a high chance of encountering another creature in the same space, whereas if it is a small noun like a fruit fly, it might take 50 or more to achieve the same high probability of bumping into and interacting with other creatures.

This kind of causal network (the DCN) is directly comparable to stock and flow diagrams from dynamic systems theory (e.g. Sterman, 2000; Jensen and Brehmer, 2003), and to the system diagrams used in complex problem solving research (e.g. Greiff and Fischer, 2013; Greiff et al., 2015): A branch of cognitive science also concerned with how people learn about complex dynamic systems, but not concerned with play. This allows for easier extrapolating of insights from research in the different fields of game-based learning, complex problem solving, system dynamics, etc. Achieving this parity would be difficult if one insisted on using traditional acyclic CBNs.
Of course, the links could represent any kind of verb, not just increasing and decreasing the populations or amounts of nouns. For example, they could represent spatial interactions, such as attraction and repulsion. Thereby, networks of nouns and verbs can be extremely powerful modelling tools. However, as discussed above regarding descriptive and prescriptive modelling notations, there is no universal solution to all modelling problems, and specific modelling notations will have to be selected to suit the context of a particular study or investigation.

DCNs can be used to model various real-world systems or games, whether it is composed of multiple objects in discrete states, or a singular object with various continuous variables (such as an aeroplane). Figure 4.3 is a simple example of a DCN.

Figure 4.3: A simple example of an aviation DCN. As before, the nodes represent nouns and their contained numbers determine the strength of their effect on other nodes, which is denoted by verbs specifying the nature of the interaction, and thus the effect on the number shown in the recipient node. Figure reproduced with permission from Tornqvist, Wen, and Tichon, 2017b, © 2017 IEEE.

4.3 Calculating Kolmogorov Complexity

A very common general measure of complexity is algorithmic, or Kolmogorov complexity. This refers to the shortest possible description of something. For example, you may tell a story to a group of friends, and each one retells it differently while still
conveying all the same information. There is theoretically a way to tell the story that minimizes the length of the story, while still containing all the same information. Kolmogorov complexity (KC) postulates that a more complex object would require a longer description than a simpler object. There are many valid and useful ways to conceptualize and measure complexity, and no doubt researchers will use those that best suit their individual context, but outlining them all is beyond the scope of this thesis. Therefore, I will discuss how KC can be applied to DCNs.

The advantages of KC include its ubiquity and extensive study as a measure of complexity. Thus, its theoretical foundations and relationship to information and Shannon entropy are well established. It is also a universal complexity measure, not bound to a particular domain such as biology or physics. Anything that can be represented as a binary string can be analysed in terms of KC, which means that causal nets that involve a different number of nodes, links, or additional components and parameters can be compared to each other, and to entirely different models such as actual simulation software, in terms of KC. This gives it a certain appeal as a universal currency of complexity.

KC was proposed by a Russian mathematician Andrey Kolmogorov in 1969 (Kolmogorov, 1969). Even though the concept is named after Kolmogorov, several other mathematicians reached similar results about the same time independently (Nannen, 2010). KC provides a way to measure the description complexity of any objects. The description complexity means how complex it is to describe a given object. KC is approximate to entropy, the quantity of information (Cover and Thomas, 2006).

For any object $x$, let $x$ be in the form of a string. Without losing generality, we suppose that every string in this section is a binary string with only “0” and “1”. Then let $U$ be a Universal Turing Machine (A Universal Turing Machine can be informally considered as any modern computers with unlimited memory).

Then the KC of $x$ regarding $U$ is defined as:
\[ K_U(x) = \min_{p: U(p) = x} l(p) \] (4.1)

where \( p \) is a program, also in binary string form, running on \( U \), which will print out \( x \) and then halt; \( l(p) \) denotes the length of \( p \). For all the programs with the same output \( x \), the KC of \( x \) regarding \( U \) is defined as the length of the shortest program, which will print out \( x \) and then halts when it runs on \( U \).

Even though the definition of KC depends on a particular Universal Turing Machine, it is universal because of the following theorem:

**Theorem 1** Let \( U \) be a Universal Turing Machine, \( \forall V \) which is another Universal Turing Machine, \( \exists c \) a constant, so for \( \forall x \in \{0, 1\}^* \) (i.e. for each binary string \( x \))

\[ K_U(x) \leq K_V(x) + c \] (4.2)

The proof can be found in (Cover and Thomas, 2006).

Due to the universality of KC, when we discuss the KC of an object \( x \), we can remove the reference machine \( U \); just denote the KC of \( x \) as \( K(x) \).

In the last few decades, KC as an objective complexity measurement has been applied in many different disciplines (Li and Vitányi, 2008) including measure the complexity of very complex and abstract systems such as the design activities for a Global Software Development (GSD) project (Wen, Kandjani, and Bernus, 2013).

Even though KC provides a way to construct an objective measurement for many different types of complex system, theoretically, KC is not computable. In a real situation, no matter how long we spend on the problem, one cannot know for sure that there is not an even shorter, more efficient program that could do the job just as well. Therefore, one can only estimate an upper bound rather than give a precise measurement of KC. In this section, I propose a scheme to estimate an upper bound of KC for a DCN.
Let $N$ be a DCN, the proposed KC measurement contains two parts: the structure section and the parameter section:

$$K(N) \leq K(N_S) + K(N_P)$$

(4.3)

where $N_S$ represents the structure information of $N$ and $N_P$ represents the parameter information of $N$.

Suppose $N$ contains $n$ nodes with $m$ directed connections. As there are possible total $n^2$ directional connections (we treat the connection from node $a$, to node $b$ and the connection from node $b$ to node $a$ as two different connections. We also consider a connection from a node to itself). Then the KC of $N_S$ is estimated as:

$$K(N_S) \leq K(n) + K(m) + K\left(\binom{n^2}{m}\right) \leq \log^* n + \log^* m + \log \frac{n^2!}{m!(n^2 - m)!} + c$$

(4.4)

where $\log n = \log_2 n$; $\log^* n = \log n + \log \log n + ...$ until the last positive term. $c$ is a constant which can be ignored when we try to compare the KC of two objects. The proof of the above formula is tedious and the detailed proof of a similar object (an $n \times n$ matrix with all elements are 0 or 1) can be found elsewhere, in (Kandjani, Wen, and Bernus, 2012).

Suppose that $N$ contains $t$ independent parameters $p_1, p_2, \ldots, p_t$ (please note we only need to count the independent parameters, some parameters can be directly calculated from other parameters; we will not include them in the KC estimation. For example, if we know the probability of raining is 0.2, we do not need to record the probability of not raining as 0.8 because it can be directly calculated as 0.8=1-0.2).

Because each parameter is associated with one entity of network elements, which are $n$ nodes and $m$ connections, we need $\log (n + m)$ bits to record the association information.
\[ K(N_P) \leq \log^* t + t \times \log (m + n) + \sum_{i=1}^{t} K(p_i) + c \quad (4.5) \]

, while \( K(p_i) \) is the KC of the parameter. It can be estimated by the rules: if \( p_i \) is an integer, then \( K(p_i) \leq \log^* + c \) (here we ignore the trivial case of \( p_i = 0 \)). If \( p_i \) has decimal places \( k \), then \( K(p_i) \leq \log^* k + \log^* \left( |p_i| \times 10^k \right) + 1 + c \). For example, if \( p_i = 0.875 \), then \( k = 3 \) and \( K(p_i) \leq \log^* 3 + \log^* (875) + 1 + c \).

Combining Eq(1,2, and 3), we have a way to estimate the upper boundary for DCNs. As a demonstration, I apply the formula to the DCNs shown in figures 4.2 and 4.3 with the following results:

Example 1. DCN in Fig.4.2

\[ n = 2, m = 2, t = 4, p_1 = 7, p_2 = 5, p_3 = 0.3, p_4 = 0.83. \] Then
\[ K(N_S) \leq \log^* 2 + \log^* 2 + \log \frac{21}{2^2 - 2} + c \leq 5 + c \]
\[ K(N_P) \leq \log^* 4 + 4 \log (4) + \log^* 7 + \log^* 5 + \log^* 3 + \log^* 83 + \log (2) + 1 + 1 + c \leq 36 + c \]
\[ K(N) \leq K(N_S) + K(N_P) \leq 41 + c(bit) \]

Example 2. DCN in Fig.4.3

\[ n = 10, m = 12, t = 9, p_1 = 0.4, \ldots p_9 = 20. \] Then
\[ K(N_S) \leq \log^* 10 + \log^* 12 + \log \frac{1021}{12(10^5 - 12)} + c \leq 63 + c \]
\[ K(N_P) \leq 124 + c \]
\[ K(N) \leq K(N_S) + K(N_P) \leq 187 + c(bits) \]

From the above examples, we find that the KC of DCN in figure 4.3 is about 4.5 times of the KC of DCN in figure 4.2.

Through such measures, one would be able to compare different studies using different games and to test hypotheses. For example, it could be used to assess
if there is an optimal level of complexity for a serious game to ensure the transfer of knowledge to the real-world domain it represents.

4.4 Conclusion

This chapter elaborated the modelling component of MMT by proposing a specific modelling method, and showing how it could be used to calculate the Kolmogorov complexity of a game. This can be applied directly in future studies, or serve as a reference point for the selection or development of similar methods of modelling for MMT in future. As previously explained, among the many modelling methods discussed in the literature, none has been established as the "correct" or universal modelling method, and thus a method must be selected or developed on a case-by-case basis to suit each individual study. This chapter can aid that process with its examples.

This forms another contribution of this research: Developing methods to objectively quantify several properties of games and their features that are relevant to outcomes of engagement and learning, suitable for use in the proposed framework and others like it. Methods to objectively quantify properties of games is instrumental to establishing connections between specific game features and specific outcomes of engagement or learning. This chapter detailed methods to quantify one form of complexity. Methods to quantify other properties of game systems also needed to be developed in the course of this research and those contributions are detailed in the coming chapters as they are relevant.

Elaborating and empirically testing every component and implication of MMT is a task too vast for this research project. Therefore, the process of Mastery was the focus of this research in the subsequent empirical experiments. The following work demonstrates the value and the manner in which MMT can be used to study more specific phenomena to do with mastery in games. Future work will have to test the other elements of Modelling and Transfer.
Chapter 5

Study 1: Evaluating Play Appeal
Factors & Engagement

The following chapter is based on a published journal paper, with some passages copied verbatim (Tornqvist and Tichon, 2019).

The first research question to which MMT was applied was RQ2: How can different combinations of game features interact to cause players to engage in off-task behaviour instead of trying to win the game? This is an important issue, as this unanticipated off-task behaviour can defeat the entire purpose of a serious game. This study will also help to demonstrate the value and the manner in which MMT can be used to design an empirical experiment to probe a specific psychological phenomenon in serious games research.

MMT emphasises the role of quantifiable properties of games to derive predictions, and has a deliberately modular structure to allow its components to be elaborated with more detailed theories and models. Therefore, a model of challenge motivation was proposed, Dynamic Probability Response (DPR), that specified what game properties can be relevant to predicting player engagement based on theories of challenge motivation. DPR specifies how to measure difficulty as a play appeal factor \((a_g)\) for use in MMT formula 3.15 \((e_p = f(a_g))\). The DPR model of challenge proposed in this research is but one example of how the modular structure of MMT
allows it to be expanded with additional details and components in future work to test more specific hypotheses while still retaining a logically coherent place in the larger picture of how educational and serious games achieve positive outcomes.

Therefore, this chapter details two significant contributions. The first is the development of the DPR model of challenge, which entails methods to objectively quantify the degree of challenge in different conditions. Methods to quantify properties of games such as these are instrumental in establishing connections between specific game properties and outcomes such as engagement and learning. As such, DPR augments the usefulness of MMT while also being critical to addressing RQ2 in the second contribution of this chapter: An empirical experiment identifying problematic combinations of game features that can cause off-task behaviour. The practical applications and theoretical significance of the experimental results are discussed.

5.1 Introduction: Motivated to Lose?

A player in *Black & White*\(^1\) stops building their city to instead pick up a boulder and send it crashing through their buildings, causing immense destruction to the player’s own hard-earned assets, much to their delight. A player in an aircraft simulator decides the instructions from the control tower are boring, and it would be more challenging to try to fly into another plane taking off, than to land safely on the runway. Why do players sometimes prefer to destroy their game world or kill their avatars? Understanding why a player may choose to ‘lose’ the game is crucial to the effective design of educational simulations and serious games, where such off-task behaviour can defeat the game’s entire purpose. In some cases, it may be a transient impulse, shortly followed by returning to pursue the game’s goal. But in other cases, the behaviour can be disruptive and frequent, defeating the entire point of the game (see e.g. Abbott, 2011). Given the fast-growing number of serious games designed to impart educational, public awareness, training and rehabilita-

\(^1\) *Black & White* is a franchise of simulation games with the player acting as a god constructing cities and conducting war. Official website can be found here: https://www.ea.com/games/black-and-white
tion outcomes, it is necessary to investigate how games are or are not successful in motivating players to interact in desired ways.

Psychological research reports the positive association of game play with emotional stability (Przybylski et al., 2012), improved relaxation and lower stress (Russoniello, O’Brien, and Parks, 2009; Snodgrass et al., 2011), reductions in depressed mood (Durkin and Barber, 2002), and positive self-esteem (e.g. Durkin and Barber, 2002; Mckenna and Bargh, 1998). However, what happens to such positive outcomes designed to be imparted by games when a player prefers to kill themselves for fun? This flaw resulting from shortfalls in game design knowledge, is also vitally important given current research which reports, when outside the game context, players closely relate themselves to their in-game character. For over a decade it has been recognised that video game exposure can lead to automatic identification with the self. That is, players associate themselves with the same traits as their character in the game. This occurs because game playing increases in the player’s memory the automatic accessibility of traits they relate to self-identity during the game (Uhlmann and Swanson, 2004; Yee and Bailenson, 2007; Yee, Bailenson, and Ducheneaut, 2009).

Skills and attitudes learned in games can transfer to the real world (for a review see Boyle, Connolly, and Hainey, 2011). For example, games that require the player to take prosocial actions can lead to increased prosocial thought (Greitemeyer and Osswald, 2011) and behaviour outside the game (Gentile et al., 2009). When intended outcomes of a game’s design are lost because players prefer to lose the game than meet the challenges they are intended to pursue, the key question of what motivates this choice and how this can be solved via game design becomes crucial to answer. This study explores what aspects of game design may derail players from pursuing the game’s intended ‘win’ goal and instead steers them to a different objective that is suboptimal or even counter-productive. It seeks to determine what aspect of game design has made ‘losing’ more fun to pursue than winning.

There is currently no model of challenge that is able to quantify the degree and type of challenge in different conditions in order to differentiate win and lose
states and make predictions of behaviour based on these. The Dynamic Probability Response model of challenge is proposed in the section below, Defining Play Appeal Factors: Proposing Dynamic Probability Response. This chapter demonstrates its use to provide these measures and predictions. This new model of challenge will enable future research on the quantification and comparative study of challenge. This forms an important initial step in objectively quantifying the motivational dimensions of games in order to answer questions about when and why one will overpower another.

The game design literature is a treasure trove of interesting theory on player behaviour, motivation and learning. Unfortunately, many remain untested, theoretical discussion. This study investigates the ‘choosing to lose’ behaviour, observed anecdotally by game designers, by testing several explanations and their hypothesised play motivations. Two of these hypothesised play motivations were derived from the game design literature and therefore currently lack direct scientific support. But each concept has indirect support from psychology studies that lend them plausibility. Studies such as the present one can help establish which game design concepts are worth developing further.

The original question behind this study encompasses a diverse set of possible situations and parameters, and therefore could be tackled in many different ways. The initial question is too broad and imprecise to be amenable to direct testing. As is often necessary in science, it needed to be reformulated into a narrower, more precise version of itself such that clear answers could be found.

One possible refinement would reform the question in terms of players breaking the rules in the process of trying to cheat, or perhaps in terms of the ‘spoil-sport’ behaviour of not caring about the game’s goals or rules at all, specifically to ruin the experience for others. Another possible reformulation is in terms of comedic transgressive play, where players might try to do the exact opposite of what the game wants in order to orchestrate ridiculous scenarios and find glitches for their own amusement. All of these approaches to the question assume deliberate transgression on the part of the player.
This study considers a best-case scenario: If players could be invested in winning the game, and trying to abide by its rules, and yet still choose to lose by a simple misinterpretation of the point of the game. According to this view, if the game designer intended X to be winning and Y to be losing, then some miscommunication might cause players to pursue Y and avoid X. This study sought to determine if this explanation could be a factor in some cases of choosing to lose, and if so, specifically which game design features might pull players astray. For example, the designer may misuse overt communicative tools (e.g. text), or more ambiguous, subtler cues (e.g. subtext) may be misinterpreted by the player.

5.2 Literature Review: Concepts of Motivation in Play

Ravaja et al. (2005; 2006) and Salminen and Ravaja (2008) measured indications of positive affect following from failure events in a game, and Abbott (2011) observed a player intentionally failing the game they had set up. Concepts discussed in the game design literature offer three possible (as yet untested) explanations: (1) the player could find the intended goal of the game boring and find ‘failure’ more challenging to achieve (implied by findings on the relative importance of challenge over other motivations. See Sherry and Lucas, 2004); (2) the game may have miscommunicated the goal of the game to the player, making them think they are supposed to pursue the failure condition (implied by Salen and Zimmerman, 2003; Schell, 2008); or (3) the failure event could result in an entertaining spectacle (like a catastrophic explosion) which exerts a motivational pull on the player (speculated by Malone, 1980; Ravaja et al., 2005).

Based on these proposed explanations, there are three possible motivators for play choice tested and compared in this project: 1) visual cues to indicate victory (game value adoption); 2) challenge; and 3) gratuitous feedback upon completion (or juice). The first hypotheses of this study are the foundational assumption that these three possible motivators have the expected effects in this study (Hypothesis 1), before then moving on to hypotheses of how they interact or interfere with each other (Hypotheses 2 and 3). These hypotheses are derived from the literature below.
5.2.1 Game Value Adoption

As discussed in the literature review section, there is a popular assumption that players need or want to be told what to do (e.g. Adams, 2009; Garris, Ahlers, and Driskell, 2002; Gee, 2005a; Juul, 2005; Portnow and Floyd, 2014; Salen and Zimmerman, 2003). Supposedly players voluntarily enter a contract of arbitrary game rules, with the expectation that those rules would lead to enjoyment. I term this concept, game value adoption. If true, then a miscommunication of the game’s goals could cause players to pursue a state other than the intended win state. For example, players may be meant to battle through the dark forest, but instead they think they are meant to follow the winding path that sends them off a cliff.

Game value adoption refers to the prediction that players will do whatever they think they are supposed to do according to the game’s rules and its definition of success or failure. This is a very straightforward prediction: players will generally try to do what they are told by the game. The deviations from this prediction are the focus of this study. But ambiguity arises when one considers how the conditions of victory or failure are communicated to players. Whatever the method of communication, overt or subtle, text or subtext, a miscommunication or misinterpretation could (according to the game value adoption concept) send players chasing a failure state instead. The question of how best to communicate through symbolism or narrative is beyond the scope of this study. Instead, I will focus on the central question of game value adoption: whether or not players do indeed prefer to pursue game states that are clearly marked as desirable (regardless of whether they actually result in any benefit or victory).

Game value adoption is an under-developed concept that raises additional questions that future studies could explore to elaborate it into a theory. It is in need of much further study and elaboration, but this is beyond the scope of this study. To investigate why players sometimes choose to lose, I simply focus on the motivational power of a communicated goal (even when achieving that goal results in no particular benefit or victory). Since it currently lacks empirical support, this will help
inform us if game value adoption is worth further investigation and development into a full psychological theory of play.

Therefore, the first hypothesis this study will test is a foundation hypothesis of whether this effect of game value adoption can be observed in the circumstances of this experiment.

H1.1: Players will pursue a win-marked game state over an otherwise identical unmarked game state (game value adoption).

5.2.2 Challenge

‘Challenge’ is a common thread found in the game design literature (see literature review above, Game Design section). Empirical research has confirmed that players play to be challenged (Alexander, Sear, and Oikonomou, 2013; Boyle et al., 2012; Hoogen et al., 2012).

Therefore, the foundation hypothesis for challenge is

H1.2: Players will pursue a challenging game state over an otherwise identical game state that is not challenging.

From the breadth of study, it is clear that challenge is regarded as one of the most important motivators of play. Players report challenge as a more important reason for play than multiple other motivations (Sherry and Lucas, 2004). Therefore, players might decide to lose the game because they are drawn to pursue something more personally challenging than the win state. The alleged primacy of challenge produces an hypothesis of how challenge might interact with other forms of play such as game value adoption.

This forms the second hypothesis, which pertains to how different combinations of game features may cause off-task player behaviour,

H2: Challenge Beats Game Value Adoption. Given a choice between two states, players will pursue the state they find more challenging even if the other
state is visually marked as the win state (derived from the high relative importance of challenge in play motivation literature, see e.g. Sherry and Lucas, 2004).

However, this requires linking broad psychological phenomena to a particular behaviour in a specific domain. Moreover, these existing theories do not adequately describe what constitutes failure states. What is required is a new model that can predict players’ tendencies for specific behaviour, based on the degree and type of challenge. For example, Nacke and Lindley (2010) had no objective means of establishing how immersive, boring, or challenging their game level designs were. A model of challenge needs to yield specific predictions about design decisions. That is, a model needs to clearly indicate to designers: ‘if I make this change to a game, I should see this behaviour from players.’ Such a model, called the Dynamic Probability Response model, is proposed and discussed in this chapter in the below section, Defining Play Appeal Factors: Proposing Dynamic Probability Response. This new model has broad potential application in quantifying challenge and agency for future studies.

5.2.3 Juice

The final game design concept examined is that players enjoy games with ‘juiciness’ (see e.g. Adams, 2009; Brown, 2015; Jonasson and Purho, 2012; Schell, 2008).

The literature reviewed in the dedicated section above, Game Design, concluded that juice could be defined as the perceptual magnitude of feedback from an action. This definition of juice produces the third foundation hypothesis,

H1.3: Players will pursue a game state with gratuitous feedback over an otherwise identical game state with basic feedback (juice).

But this study is interested in juice that is tied to the advent of a game state, such as winning or losing. If the advent of a particular game state results in a delightful spectacle, like a catastrophic explosion, then it would be juicy and could tug players away from pursuing the win state. For example, Malone (1980) speculates that,
One potential problem with... a fantasy catastrophe (like a person being hung) is that the catastrophe may be so interesting that players try to get wrong answers so that they can see it. (p. 67) [emphasis in the original]

This speculation is echoed by Ravaja et al. (2005) in their explanation for finding evidence of positive affect following from a failure event. Note that, while there is definitely an argument to be made that challenge could overpower juice (such an hypothesis would be worth exploring), the citations in the above section on challenge do not explicitly claim or predict this. In contrast, both of these sources on juice do make explicit reference to juice overpowering in a context in which to win would be challenging, and is clearly indicated as the goal of the game (game value adoption), producing the second hypothesis of interaction,

H3: Juice Reigns Supreme. Players will pursue a juicy game state over a challenging and/or win-marked game state (derived from the speculations of Malone [1980], and Ravaja, et al. [2005]).

The above literature and derived hypotheses (e.g. see the above quotation) imply a simple hierarchical theory of motivation interaction, where one form of play at higher level in the hierarchy overpowers the one/s below it. This is one of the simplest possible interactions to posit, and therefore will be used as the starting point in this study. The observations of Malone (1980) and Ravaja et al. (2005) implied that juice overpowers both challenge and the visually indicated win state of the game, putting it above both of those in the hierarchy (H3). Challenge was hypothesised to be next down in the structure, above win-marking but below juice (H2). This was because the above authors explicitly placed juice above challenge, but aside from this, challenge is often assigned high relative importance in play motivation literature (see e.g. Sherry and Lucas, 2004). This left win-marking (game value adoption) at the bottom of the hierarchy.

Placing juice at the top of this hierarchy (H3) means that it overpowers both items below it in the hierarchy. This carries with it the logical consequence that it will overpower them not only when they are together, but also in isolation,
H3.1 Players will pursue a juicy game state over a win-marked game state.

H3.2 Players will pursue a juicy game state over a challenging game state.

H3.3 Players will pursue a juicy game state over a challenging win-marked game state.

This study first tested if the hypothesised motivational pull of juice could be found at all (H1.3), and then tested the hypothesis that it can overpower other motivators such as challenge (H3.1, 3.2, and 3.3). Generating such data will help determine if the concepts of juice and game value adoption are worth developing into theories, and (if so) to also produce data to lay some of the preliminary groundwork for transitioning the concepts from vague game design truisms to rigorous scientific models of motivation and behaviour.

Through the hypotheses derived and detailed above, this study first tested the foundational hypothesis (H1) that these three concepts of motivation could affect player behaviour, and then investigated RQ2 by testing the implied hierarchical theory of motivation interaction in the literature (H2 and 3). These hypotheses, if supported, would provide an explanatory mechanism for off-task behaviour in a game if the game had a flawed design that combined features in such a way as to distract players from the goal (e.g. a juicy fail state, or a challenging fail state). This study therefore tests a hierarchy with juice at the top, challenge in the middle, and game value adoption at the bottom.

5.3 Study Framework & Definitions

5.3.1 Different Win-Lose Definitions

What does it mean to pursue failure in a game, and what does it mean to try to win? How can one precisely define what constitutes winning or losing? One way to define the win state is as whatever the game designer intended to be the win state.
In which case, the only way to definitely answer the question is to simply ask the designer (this is the designer’s win-lose definition).

Another way to try to objectively define the win state is to use some criteria pertaining to information found solely in the game itself (the game’s win-lose definition). That method would sidestep the need of asking the designer. Exactly what criteria should be used to objectively determine a game’s win state is debatable. In clear cases, there could be actual ‘Game Over’ or ‘Victory’ screens. In other cases, it might be less literal, with the win state resulting in celebratory fireworks or a tense chase sequence followed by a climactic explosion, providing narrative closure and release of tension. Win-lose states might be communicated with certain carrots or sticks to provide incentives or threats that hopefully indicate the intended win or lose state. Notice that in these examples, game design concepts like game value adoption, narrative, juice, and challenge are used to try to tug players towards the designer’s intended win state and away from their intended lose state. Therefore, to look for objective definitions of win-lose states in the game itself will require an understanding of how such game design concepts can be used to serve as carrots or sticks to communicate desirable and undesirable states to players.

Thirdly, the win and lose states could be defined in terms of the player’s interpretation (the player’s win-lose definition). In which case, a win state (or lose state) is whatever the player perceives to be a win state (or lose state). By this definition, the only way for a player to pursue a lose state is to deliberately transgress. Using the player’s win-lose definition, it would be impossible for a player to lose in the way I hypothesised: By misinterpretation of the game’s win state (because the win state is whatever they interpret it to be). For that reason, this study does not adopt this player-based definition of win-lose states for this study.

Note that each of these three definitions are likely to be causally related: The designer has an intended win state, and so designs the game with carrots and sticks to communicate that win state, and players feel which way the game is pulling or pushing them to form an interpretation of what the win state is supposed to be. If there is a problem at any step in this process, then it is possible that the player’s
interacted win-lose states will not match the designer’s intended win-lose states. That is the scenario I assume for this study, and around which the experiment was built.

The experiment was designed to engineer conditions where the game’s win-lose definition was clear and concordant according to the pushes and pulls found in the game, and then compare these to conditions where the game’s win-lose definition was ambiguous due to conflicting pushes and pulls. The player’s win-lose definition was operationalised according to their behaviour: which state they chose to pursue or avoid. The extent to which a difference in design results in a play motivation pulling the player away from the original win state and towards the other state (i.e. results in a change in the player’s operational win-lose definition) will be termed a lose-tug effect.

In this study, there were only two choices presented simultaneously. Therefore, lose-tug effects were conceptualised as how much the player invested in pursuing the non-win state relative to the win state, and how much this value changed across conditions when, for example, juice was introduced to one of the states. See below sections, Measures and Results, for details on how these values were measured and calculated.

The designer’s win-lose definition, and how it fails to be properly translated into the game’s win-lose definition, is outside the scope of this study. I simply sought to engineer the general circumstances of such a game design flaw, and not to established how or why that step in the process can go wrong. To minimise bias due to preconceptions, I have tried to maintain a neutral stance toward which state is actually the ‘win’ state, so as to not influence players in the experiment and just observe which states they felt were most appropriate to pursue. This approach was thought to better capture the idea of players misinterpreting of the goals of the game. This study refined the broad question, ‘why do players choose to lose’, into a more precise form, ‘can a difference in a game design result in conflicting play motivations that pull players away from the win state? If so, which play motivations win out when they conflict?’
5.3.2 Defining Play Appeal Factors: Proposing Dynamic Probability Response

Each of these three motivational concepts needs to describe both win and failure states – states that attract the player to complete them, and states that make the player avoid them. This will be referred to as the valorisation of the game state: whether the player regards it as good or bad. To valorise is to assign value or merit to something. If the player does not assign any value (positive or negative) to a system state, then they cannot be expected to pursue or avoid that state, and therefore we would only predict player behaviour to be affected by valorised states (either positive valorisation for win conditions, or negative valorisation for failure conditions). Design features that valorise a game state (positively or negatively) may indicate aspects of motivational psychology that can be put to use in future game design, whereas design features that do not valorise game states (positively or negatively) may bring into question the underlying theory of play on which those designs were based.

According to the juice concept, a win state (positive valorisation) will be juicy in terms of its feedback to the player, and a fail state will be inert. Specifically, the feedback will be juicier if it is of a greater perceptible magnitude (e.g. greater salience, complexity, etc.), without conveying any additional relevant information than the purely informational feedback of the non-juicy condition. According to game value adoption, win and fail states are defined by the game’s rules: players will generally pursue whatever the game says is good and avoid whatever the game says is bad. Specifically, if a state is visually indicated as good or desirable, then players should be seen to pursue it (regardless of whether it actually yields any benefit or victory). Finally, there is the motivation of challenge in and of itself. To undertake the current study, a definition of challenge’s win and fail states needed to be determined. This was achieved by proposing the Dynamic Probability Response (DPR) model of challenge.

Psychological phenomena such as the effort heuristic, the mastery hypothesis, and flow predict that people will have an inherent interest in that which is
challenging. But to study gameplay scientifically, a formal model for quantifying the amount and type of challenge in a game system will be necessary.

Most theories concentrate on explaining motivation for achievement rather than aversion to failure. Those that attempt to explain failure define it in terms of that which is too easy to achieve (see e.g. Weibell, 2011). However, it is unclear if easy tasks are likely to result in negative affect and boredom, or positive affect and relaxation (Engeser and Rheinberg, 2008). Some theories specify circumstances in which people should feel a sense of failure, or try to avoid something (see e.g. Aponte, Levieux, and Natkin, 2011; Elliot and Harackiewicz, 1996; Engeser and Rheinberg, 2008; Hoska, 1993; McMillan, 2013; Weibell, 2011; Fraser, Katchabaw, and Mercer, 2014), but these are in response to not achieving a state, or explained in terms of the characteristics of the individual rather than in terms of differences in the game states (e.g. flow theory describes boredom as when the individual’s skill is far beyond the difficulty of the task). Because this study seeks to develop knowledge useful for designing games and simulations, it is concerned with a change that can be made to the game which results in a sense of failure about the advent of a particular game state. Thus, the model used is one that focuses on when players do achieve a state rather than when they do not, and on variations in the game’s design rather than in the player’s skill.

Across research in both psychology and game design, challenge is described in terms of a one-dimensional continuum from easy to hard (e.g. Atkinson & Feather, “the incentive value of failure is more negative the easier the task” p. 331, 1966, as cited in Weibell, 2011). But a two-dimensional model of challenge can distinguish between positive and negative game state valorisation. Consistent with the premise of the game design and psychological theories above, one axis of difficulty-to-achieve should correspond to feelings of success when that game state is reached, and the other axis of difficulty-to-avoid should correspond to feelings of failure.

This creates a two-dimensional space (a 2D Difficulty Space or 2DDS), and any given game state must occupy a specific point on that space, as in figure 5.1. It
may shift across this space as strategic variables evolve during play, but at any one point in time, it can only occupy one point on the space.

Figure 5.1: The 2D Difficulty Space (2DDS) of the Dynamic Probability Response model which is used to describe challenge in this study. Figure adapted from Tornqvist and Tichon, 2019.

Specifically, Dynamic Probability Response defines challenge as the probability of a state change, given the player’s efforts for or against that change. When that new state is highly likely to occur despite the player’s attempts to prevent it, then it is difficult to avoid. An example of this kind of challenge is trying to outrun a lion - Being eaten by the lion is difficult to avoid (figure 5.1, point D). When the new state is very unlikely to occur despite the player’s attempts to pursue it, then it is difficult to achieve. For example, landing an arrow dead on the bullseye would be difficult to achieve (figure 5.1, point A). These opposite quadrants (top left and bottom right inside the graph, in figure 5.1, points A and D) of the 2DDS valorise the state oppositely. In contrast, the centre line from bottom left to top right should remain neutral, with no valorisation (figure 5.1, points B and C).

Just as important as what changes will affect valorisation are what changes will not. As such, this model also describes another commonly-discussed game element: agency. Here agency refers to the power of player actions to affect the game world. Agency is related to the motivational concept of autonomy in Self-Determination Theory (Deci and Ryan, 2000; Ryan, Rigby, and Przybylski, 2006). Adjustments by a game designer that change a game state to move that state diago-
nally along the 2DDS between bottom left and top right will change agency without affecting valorisation. For example, if the probability of the new state is entirely dependent on the player’s efforts, then it is trivial, and the player has full agency with respect to that state change. For example, toggling between crouch and stand by pressing a button on a gamepad is a trivial state change with no valorisation (figure 5.1, point C). On the other hand, if the probability of the new state is uncorrelated with the player’s efforts, then the player has no agency regarding that state change. For example, if "winning" is simply whether or not "heads" comes up when the referee flips a coin, then the event has no inherent valorisation related to challenge\(^2\) and the player has no agency with respect to that outcome (figure 5.1, point B). That this model provides a description of agency is incidental to the study at hand, but has important implications for future games research about agency and control (e.g. Rogers, Dillman Carpentier, and Barnard, 2016) and how it relates to challenge.

Player effort can be plotted on one axis, and probability of the game state in question occurring can be plotted on the other axis, to gain more information. The x axis can represent player effort, where a value of one on the x axis represents the player attempting to achieve the game state in question; a value of zero on the x axis of player effort represents the player trying to avoid the game state in question; and a value of 0.5 on the x axis represents that the player is neither trying to achieve nor avoid the game state in question. Probability of the advent of the game state can be plotted on the y axis, and combining this with the x axis of player effort produces a challenge curve (figure 5.2).

\(^2\)Note that DPR as a model of challenge does not attempt to explain the phenomenon of a gambler experiencing a thrill on correctly guessing the outcome of a coin flip. That would require a different psychological theory to explain.
Figure 5.2: Some example Challenge Curves elaborating certain points on the 2D Difficulty Space. Each equivalent 2DDS point is shown to the right of each Challenge Curve. Refer to figure 5.1 for details. Figure adapted from Tornqvist and Tichon, 2019.
A challenge curve provides a visual representation of challenge according to the logic of Dynamic Probability Response (another representation being the 2D difficulty space of figure 5.1). The two ways are largely interchangeable but there is one point of useful difference. Challenge curves can represent an additional layer of detail that the 2DDS cannot capture. For example, a challenge curve may be exponential, parabolic or some other shape, which would have no equivalent on the 2DDS (see curves 7 and 8 in figure 5.1, both showing the same difficulty-to-achieve but by differently-shaped curves). As a result, the 2D difficulty point best describes a state broadly, and challenge curves capture more subtleties.

To apply DPR in a study to objectively quantify challenge requires mathematics. For example, if a study was able to measure the probability of a particular game state under the different conditions of A) the player trying to avoid that state, B) the player not trying to avoid or achieve that state, and C) the player trying to achieve that state, then these data points could be used to construct a challenge curve for that game state (with the three conditions A, B, and C plotted on the x axis, and probability of the game state on the y axis, as in figure 5.2). Once these measurements have been obtained and plotted, then a 2DDS point can be extracted from the challenge curve, and a 2DDS point can be used to make predictions of future player behaviour (this process is carried out in detail in the below section, Checking the Game’s Challenge Curves).

To extract a 2DDS point from a challenge curve, let \( dc \) be difficulty-to-achieve, \( dv \) be difficulty-to-avoid, \( g \) be the gradient (slope) of the challenge curve, and \( \bar{P} \) be the mean probability of the challenge curve (i.e. average value of the y axis):

\[
(\bar{P} \geq 0.5 \rightarrow dc = 1 - g - 2(\bar{P} - 0.5)) \land (\bar{P} < 0.5 \rightarrow dc = dv - 2(\bar{P} - 0.5)) \quad (5.1)
\]

\[
(\bar{P} \geq 0.5 \rightarrow dv = dc + 2(\bar{P} - 0.5)) \land (\bar{P} < 0.5 \rightarrow dv = 1 - g - 2(\bar{P} - 0.5)) \quad (5.2)
\]
More complex formulas may be useful to extract a more accurate 2DDS point from complex challenge curves that deviate significantly from a linear line. For example, there is an argument to be made that a sharp spike in the challenge curve should be weighted more heavily than a smooth slope for psychological reasons, but that discussion is not necessary for this study. These basic formulas will be sufficient for this study and for many simple cases, where the challenge curve is linear or close to linear.

Once a 2DDS point has been calculated for a game state, two variables can be extracted that can serve as play appeal factors \((a_g)\) in MMT formula 3.15 \((\tilde{e}_p = f(a_g))\). The first and most relevant to this study is the valorisation due to challenge. Let \(v\) be valorisation, \(dc\) be difficulty-to-achieve, and \(dv\) be difficulty-to-avoid:

\[
v = dc - dv
\]  

(5.3)

A negative \(v\) would represent a "failure" state in terms of the game’s win-lose definition. This negative valorisation would therefore predict that (all else being equal), the player’s win-lose definition would also designate the game state as a "failure" state, and therefore result in player behaviour that entails trying to avoid that game state occurring. On the other hand, if \(v\) is positive, then this would represent a "win" state in the game’s win-lose definition, which (all else being equal) should influence the player’s win-lose definition and result in player behaviour that involves trying to achieve that game state.

The other possible play appeal variable that can be extracted from a 2DDS point is a measure of agency. Let \(\alpha\) be agency, \(dc\) be difficulty-to-achieve, and \(dv\) be difficulty-to-avoid:

\[
\alpha = 1 - \max\{dc, dv\}
\]  

(5.4)

This would produce a measure of agency in terms of how much influence the player has over an outcome. Agency has often been considered a play appeal
factor (e.g. Rogers, Dillman Carpentier, and Barnard, 2016), but it is not the subject of this study. Therefore, while it won’t feature in this particular study, it is a potentially useful feature of DPR that could prove useful in future work that employs DPR. For example, some theories of challenge such as flow (Czikszentmihalyi, 1990) posit that engagement will be low not only when challenge is low, but also when challenge is too high if the player’s skill is too low to cope with the challenge. In which case, a high \( v \) may also need to be coupled with a moderately high \( \alpha \) in order to result in high engagement with the game. Such a theory could take \( v \) and \( \alpha \) as play appeal factors \((a_g)\) in MMT formula 3.15 \((\hat{e}_p = f(a_g))\) to produce a function that predicts engagement based on these objective measures of game properties \((\hat{e}_p = f(v, \alpha))\). While this is outside the scope of this study, it helps demonstrate the usefulness of DPR and similar methods to objectively quantify game properties to further games research methodologies.

Given that the purpose of DPR is to quantify the challenge present in different game conditions for the purposes of conducting psychology studies, it is much more important to be able to convert from a measured challenge curve into a 2DDS point, than to convert in the other direction, from a 2DDS point to a challenge curve. As explained above, a 2DDS point could map to multiple challenge curves and so there is no definitive “correct” way to convert from a 2DDS point to a challenge curve (there is simply insufficient information to describe the curve). But if a basic curve would be useful for a study, then the below formulas can provide a simple, linear challenge curve. Let \( dc \) be difficulty-to-achieve, \( dv \) be difficulty-to-avoid, \( g \) be the gradient (slope) of the challenge curve, and \( \bar{P} \) be the mean probability of the challenge curve (i.e. average value of the y axis, which would be the centre point on a linear curve):

\[
\bar{P} = 0.5 + \frac{dv - dc}{2} \tag{5.5}
\]

\[
g = 1 - \max\{dv, dc\} \tag{5.6}
\]
Dynamic Probability Response is intended to address the question of why players choose to lose by predicting which state players will choose to pursue. Note that this model entails the following assumptions and constraints:

Because difficulty-to-achieve valorises positively and difficulty-to-avoid valorises negatively, a state composed of equal quantities of both is therefore neutral (e.g. curve 1, 2, or 5, in figure 5.2). For example, it is equally difficult to make it rain as it is to prevent it from raining (curve 2), and so when it rains one would not feel a sense of success or competence. As Togelius and Schmidhuber (2008) point out, "a game is no fun if it can be won by doing nothing at all or acting randomly" (para. 24).

DPR implies that it is impossible to increase the difficulty of a game state (in either dimension) without also reducing player agency with respect to that state (e.g. compare curve 1 to curve 6 or 7). In other words, agency is sacrificed to create either kind of difficulty (see formulas 5.3 and 5.4).

DPR does not describe the conditions for how challenge can be too great as to turn into frustration. To achieve that end, one would have to extend DPR by combining it with a model that does describe this phenomenon, such as flow (Czikszentmihalyi, 1990).

DPR allows quantitative description and comparison of the patterns that game designers notice anecdotally, as demonstrated in the later section, Checking the Game’s Challenge Curves. Crucially, it also makes predictions about how a designer can change a game system and expect to see specific player behaviour as a result.

In this experiment, only two particular challenge curves will be examined. But Dynamic Probability Response is much more general and versatile, which will allow knowledge generated by this study to inform future studies that look at very different questions via this model.
### 5.4 Method

#### 5.4.1 Game

For this study, it was necessary to design and build a simple game with adjustable parameters (e.g. difficulties, juiciness, and which game state is visually shown as a win state) to test hypotheses.

The game (figure 5.3) is a simple chase game, where the player uses a joystick to control an arrow as their token (or *avatar*), moving it across the 2D environment to try to chase down, or run away from, autonomous game tokens (or *agents*) that move according to some basic rules. The environment is a simple square that is flat (i.e. no obstacles or hills) and wraps around: if an agent moves past the left border of the screen, it appears on the right-hand side (and the same occurs for the top and bottom). Each condition started with the player in the centre and the agents placed randomly.

The game state examined was whether or not the avatar touches a given agent. Difficulty-to-achieve was generated by making the agent run away from the avatar, and difficulty-to-avoid by making the agent chase the avatar.
5.4.2 Measure

The data for this study was recorded in-game. The amount of free time voluntarily spent on an activity is often used as a measure of intrinsic motivation (Boggiano and Ruble, 1982; Cameron, Banko, and Pierce, 2001; Elliot and Harackiewicz, 1996; Lieberoth, 2015; Ryan and Deci, 2000). It is often treated as an indicator of liking a game (Johnson et al., 2014; Malone, 1980). Therefore, the measure in this study was based on this premise and extended to not just give an absolute measure of time for one option, but a relative, continuous, and normalised measure to capture the extent to which one option vs. the other was pursued over time. The measure itself was derived from studies measuring control skill in complex systems as the Euclidean distance between a user’s input and the ideal possible input to guide the system to a particular outcome (e.g. Beckmann and Goode, 2014; Goode and Beckmann, 2010; Goode, 2011). As a measure of player interest in pursuing either state, the game calculated (for each agent) the ideal maximum velocity the player’s avatar would have if he was trying to pursue the agent, and then compared this with the avatar’s actual current velocity (the player controls the avatar’s velocity one-to-one via the joystick).
The difference between these two vectors (calculated as the distance between them, as though they were locations in space rather than hypothetical velocities) was the measure of player interest (see figure 5.4). There are myriad possible mathematical transformations that could be done to weight, for example, smaller differences or larger differences. For example, in clustering analysis it is justifiable to use square Euclidean distance (Wang et al., 2008). However, this is not a cluster analysis, and without a compelling reason to think that a particular transformation accurately captures a specific and significant component of the human preference between game states in this kind of situation, no additional mathematical steps were performed.

Figure 5.4: An example showing the vectors used to calculate player preference between pursuing different agents. Figure adapted from Tornqvist and Tichon, 2019.

The minimum of zero indicated that the avatar was moving towards that agent as fast as possible. The maximum of one indicated that the avatar was moving away from that agent as fast as possible. A reading of .5 indicated that the avatar was not moving towards or away from that agent.

Therefore, if a player preferred to chase after agent number one, the measurement would be less than .5 for agent number one (e.g. .34 < .5), and that mea-
measurement should be less than the measurement for agent number two (e.g. \(0.34 < 0.46\))

At the conclusion of each game session, the program automatically took all the frame-by-frame readings and averaged them into the one measure for the entire game session. The measure used for this study was the average of frame-by-frame recordings of the distance between an ideal avatar velocity and the actual avatar velocity, over an entire game session.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean Relative Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>A participant’s preference to pursue agent 1 over agent 2 in a particular condition</td>
</tr>
<tr>
<td>Measurement</td>
<td>Preference = Euclidean distance between player’s current velocity and the velocity appropriate to intercept agent 1, at time (t), normalised (value of 0 indicates complete preference toward agent, value of 0.5 indicates indifference, value of 1 indicates complete aversion away from agent)</td>
</tr>
<tr>
<td>Measurement Processing Step 1</td>
<td>Relative Preference = Preference measurement for agent 1 minus preference for agent 2, at the same time (t), normalised (value of 0 indicates complete preference for agent 1, value of 0.5 indicates no preference between agents, value of 1 indicates complete preference for agent 2)</td>
</tr>
<tr>
<td>Measurement Processing Step 2</td>
<td>Mean Relative Preference = Average relative preference measurements across all time (t) for a participant in a condition</td>
</tr>
<tr>
<td>References</td>
<td>Boggiano and Ruble, 1982; Cameron, Banko, and Pierce, 2001; Elliot and Harackiewicz, 1996; Lieberoth, 2015; Ryan and Deci, 2000; Johnson et al., 2014; Malone, 1980; Beckmann and Goode, 2014; Goode and Beckmann, 2010; Goode, 2011</td>
</tr>
</tbody>
</table>

Table 5.1: Summary of the measure used to test hypotheses in this study.

Qualitative pilot experiments were also used to get an approximate value for the amount of time needed for a game session, and thus each condition. If the time was too short, the measure used to approximate player interest in each agent may not have had enough data for one agent to clearly be preferred over the other. But if it was too long, there was the risk that the player would begin to feel bored.

---

3This is not a perfect measure. For example, if the avatar was not aiming directly at an agent, but instead trying to predict where it would be to intercept it, then this measure would read less than maximum because the avatar was not moving directly towards the agent. But it was considered a superior and more objective measure than, for example, getting the participants to complete a questionnaire about which state they preferred to pursue.
and stop playing. For this study, conditions automatically ceased after 15 seconds had elapsed; then the next condition would start.

5.4.3 Checking the Game’s Challenge Curves

Where Nacke and Lindley (2010) had no objective way to establish whether or not they had succeeded in creating boring or challenging designs for their experiments, Dynamic Probability Response allows researchers to do just that.

As the game was designed to maximise its utility for future study, the difficulty parameters stretch from trivial (curve 1) to impossible on both axes (i.e. impossible for difficult-to-achieve – curve 4 – and inevitable for difficult-to-avoid – curve 3). In this experiment, the difficulty was never set to one (which would be virtually impossible/inevitable) because some psychological and game design theories (e.g. flow) suggest that too much difficulty will have a similar deleterious effect (e.g. frustration) on player interest as too little difficulty (e.g. boredom, as mentioned in Przybylski, Rigby, and Ryan, 2010; Nacke and Lindley, 2010). The aim was not to discover the boundaries that demarcate frustration from flow. Therefore, various values for the 2DDS settings in the game were trialled in qualitative pilot experiments, in order to hone in on the values at which the difficulty would minimise the chances of frustration or boredom affecting the results. Based on these pilot experiments, the difficulty settings used were .6 difficult-to-achieve vs. .55 difficult-to-avoid.

To check if the game actually had the difficulty dynamics necessary for these experiments, as commonly done in game research (see e.g. Togelius and Schmidhuber, 2008), player behaviour was simulated by controlling the avatar with a simple artificial intelligence (AI) that could be directed to pursue or avoid specific agents.

To get three points on the player effort axis of the challenge curve, a specific agent was selected, then the AI avatar was directed to either chase after it (player

---

4Note that this AI is not intended to be a perfect simulation of player behaviour. It is a simulated ideal player that has omniscience, perfect reflexes, and unwavering interest in whatever it is told to do. This keeps constant the variables of motivation and skill in determining how likely the player is to make contact with the agents. The purpose of this AI was simply to measure the actual difficulty generated by different parameters.
effort = 1), run away from it (player effort = 0), or just choose random destinations to run to (player effort = .5). Moving to a random location, then choosing a new random location to move to, ad infinitum, was the preferred method of approximating .5 player effort because it looked visually more like player behaviour than either not moving at all or choosing a random velocity every frame.

The average distance (over the course of the game session) between the avatar and the agent was the measure of the chance of the state occurring. It was assumed that the closer the agent was on average, the more likely the avatar was to touch it.

Data was gathered from a sample of 20 automated game sessions (with the avatar controlled by the simple AI) per difficulty setting, per player effort level (figure 5.5).

Figure 5.5: The challenge curves measured from the difficulty settings. The settings used in this study are confirmed to be within the range needed. Figure adapted from Tornqvist and Tichon, 2019.

Points on the 2D difficulty space can be calculated from these challenge curves. The challenge curves were normalised to turn proximity into a probability of making contact between zero and one, using zero distance as the maximum likelihood of contact, and the mean proximity of the zero Trivial setting (zero difficulty-to-achieve and zero difficulty-to-avoid) at zero player effort (i.e. maximum evasion) as the assumed minimum likelihood. Then 2D difficulty points were calculated (see
STUDY 1: EVALUATING PLAY APPEAL FACTORS & ENGAGEMENT

formulas 5.1 and 5.2). The resulting graph confirms that these difficulty settings can be used for this experiment, but it also reveals a distortion in the difficulty space\(^5\).

The resulting graph (figure 5.6) shows the two points compared do fall on either side of the mid-line, meaning they should valorise the game state in opposite ways - one positively and one negatively. This confirms that differing types of challenge were compared in this study, and demonstrates using Dynamic Probability Response to measure the degree and type of difficulty.

![Figure 5.6: The 2D difficulty points measured from the difficulty settings. This confirms that the two settings lie on opposite sides of the measured midline and should therefore differ in their valorisation of the game state. Figure adapted from Tornqvist and Tichon, 2019.](image)

5.4.3.1 Implementing the Parameters in the Game

The possible instantiations of each of the win conditions is prohibitively vast (see section below, Limitations and Future Work). There are innumerable ways to create a challenge, to create juice, or to communicate the goal. This study tested one instantiation of each of these. Challenge was generated with a chase mechanic.

For the purposes of this research, players were told which game state was the ‘good’ state to pursue via leveraging conventions of visual symbolism in games. The symbol of a heart, commonly used to communicate potential positive gains such

\(^5\)The measured mid line is lower than \(x=y\) and bent, and similar difficulty settings are not mirrored around the mid line. This could be due to having used proximity as a stand-in measure for probability, or due to an asymmetry in approach and avoidance behaviour (as found in e.g. Buchan et al., 2005; Kahneman and Tversky, 1984).
as extra lives, was used to signify the desirability of a state. If players are eager to adopt the game’s definitions of desirable and undesirable states, then they should pursue the heart-marked state. The not-win agent was a white square.

A white ring encircled agents that were juicy. This white ring indicator was counterbalanced, with half of the participants getting the ring to indicate juiciness, and the other half getting the ring to indicate the juiceless agents. This visual marker was to help players keep track of and chase a singular juicy agent among other fast and erratically moving agents. Acclimatising sessions ensured players would learn whether the white ring indicated juiciness or juicelessness.

In this study, juice was achieved by a simultaneous combination of sound effects, camera shake, object wobble (in translation and scale), particle effects, and slowly fading decals on the environment (see figure 5.7).

![Figure 5.7: Screenshot showing the visual effects used to create juiciness (splash particles and decals) for this study. Figure adapted from Tornqvist and Tichon, 2019.](image-url)

The juice setting controls all elements of the juice effects simultaneously along a continuous scale. For the juicy conditions, it was set to one (maximum). For all other (non-juicy) conditions, it was set to .1 so that players did receive some auditory and visual feedback confirming they had made contact. At this lower setting, the audio feedback sounded like little more than a click, the decal was smaller than the icons representing the agents, shaking was only slight, and the particle effects were not visible. If there were no audio or visual indicators of when contact was made, then any difference between the non-juicy and the juicy conditions
could just be due to providing any feedback at all, and not specifically the magnitude of the spectacle. This way, an unobtrusive form of feedback was compared to a gratuitously spectacular one. This would meet juice’s definition of conveying no additional relevant game information than the non-juicy condition, but instead including additional non-informative elements to increase the perceived magnitude (e.g. salience, complexity) of the feedback.

5.4.4 Participants

A laptop PC with an Xbox 360 gamepad was set up just outside a cafe situated next to a university in Brisbane, Australia (Ethics information in appendix C). A sign saying “Free Science” was used to attract potential walk-by participants, but the vast majority were directly approached and asked if they would voluntarily participate in a psychological experiment by playing a game. Over seven days, a total of 67 participants were recruited. To increase the number of samples that could be gathered in the available time to increase statistical power no demographic data was collected in this experiment.

5.4.5 Procedure

Participants were told what the controls of the game were. They were also advised the different conditions were entirely independent of each other and unaffected by what the player did in previous conditions. This was to minimise any attempts at meta-gaming.

There were nine conditions (see figure 5.8) that all participants experienced. To ensure the results weren’t just a result of all participants experiencing these conditions in the same order, the order was counterbalanced by an incomplete balanced Latin Square design. Counterbalancing by randomly reordering the conditions was considered, but that could cause a disproportionately large number of participants to experience very similar orderings just due to random chance. More systematic methods of reordering like Latin Square avoid this problem.
Figure 5.8: The nine conditions and seven comparisons to test hypotheses. H1.1- Players will prefer the win-marked agent; H1.2- Players will prefer to chase the difficult-to-achieve agent; H1.3- Players will prefer the juice; H2- Challenge will tug from the win-marked agent; H3.1, 3.2, & 3.3- Juice will tug from both challenge and win-marking, and the combination of the two. Figure adapted from Tornqvist and Tichon, 2019.

The first two conditions for all of the participants were acclimatising sessions (when added to the nine data-gathering conditions, there were 11 conditions in total, and two minutes and 45 seconds of gameplay). The first one was to ensure that players understood the white ring indicator of juiciness. It involved one normal agent, one juicy agent, one win-marked agent, and one win-marked juicy agent. All four agents were immobile for this condition to allow the player to touch them at their leisure and see the results. The second acclimatising condition was to allow the players to explore the dynamics of the game and grow familiar with them before the first experiments. This was to minimise the effects of exploratory play on the results. This acclimatising session had one of each kind of agent that the player would see in the game, all present at the same time.
5.5 Results

Descriptive results are summarised in table 5.2. The tests generated two values: one player preference measurement for each of the agents present in each condition. To get a measure of the player’s relative preference between the agents, these two measurements were subtracted from each other (see table 5.1), which resulted in the measurement ranging from negative one (indicating that the player unambiguously preferred the first agent with no interest in the second) to positive one (indicating that the player unambiguously preferred the second agent with no interest in the first). Zero indicated a complete lack of preference, and as expected, the relative measurement of the control condition (where both agents were indistinguishable) was very close to zero (-0.01), and provides an indication of the margin of error to expect in this data. As can be seen in figure 5.9, the only other condition that came close was win-marking against juice (-0.04), suggesting that juice and game value adoption might be fairly evenly matched in terms of their motivational pull, which is further supported by the closeness of the juice condition (-0.10) to the win-marking condition (-0.10).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.12</td>
<td>-0.25</td>
<td>0.31</td>
</tr>
<tr>
<td>2. Win-Marking</td>
<td>-0.10</td>
<td>-0.07</td>
<td>0.22</td>
<td>-0.51</td>
<td>0.52</td>
</tr>
<tr>
<td>3. Juice</td>
<td>-0.10</td>
<td>-0.07</td>
<td>0.20</td>
<td>-0.59</td>
<td>0.34</td>
</tr>
<tr>
<td>4. Challenge</td>
<td>-0.18</td>
<td>-0.18</td>
<td>0.12</td>
<td>-0.45</td>
<td>0.07</td>
</tr>
<tr>
<td>5. Win-Marking + Challenge</td>
<td>-0.32</td>
<td>-0.32</td>
<td>0.17</td>
<td>-0.65</td>
<td>0.05</td>
</tr>
<tr>
<td>6. Win-Marking vs. Challenge</td>
<td>0.15</td>
<td>0.12</td>
<td>0.16</td>
<td>-0.14</td>
<td>0.52</td>
</tr>
<tr>
<td>7. Win-Marking vs. Juice</td>
<td>-0.04</td>
<td>-0.07</td>
<td>0.26</td>
<td>-0.46</td>
<td>0.53</td>
</tr>
<tr>
<td>8. Challenge vs. Juice</td>
<td>-0.17</td>
<td>-0.18</td>
<td>0.13</td>
<td>-0.54</td>
<td>0.10</td>
</tr>
<tr>
<td>9. Both vs. Juice</td>
<td>-0.23</td>
<td>-0.20</td>
<td>0.17</td>
<td>-0.53</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 5.2: Descriptive Statistics across all conditions. The measure is mean relative preference calculated based on relative velocity as defined in the above section Measure, table 5.1. Figure adapted from Tornqvist and Tichon, 2019.
By comparing the conditions and subtracting these mean relative preferences (that is, looking at the size of the difference between the bars in figure 5.9), one can see how changing one characteristic of one of the two agents affected mean relative preference between them. Thus, the difference of mean relative preference gives a measure of a ‘lose-tug effect’ – a change in how much players preferred each agent when one feature was different. The differences of means in figure 5.10 reveal that all the effects are in the direction predicted by the hypotheses.
Figure 5.10: The lose-tug effects (difference in mean relative preference) found from the planned comparisons, giving an impression of relative effect sizes. In each column are two agent icons overlapping to indicate that this was the agent whose property changed, and what it was changed to (thus indicating for each column in figure 5.10, which two bars of figure 5.9 were subtracted to find the difference in mean relative preference). This is also reflected in the text box above each column. Figure adapted from Tornqvist and Tichon, 2019.
<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Comparison</th>
<th>Mean Difference (I-J)</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1.1</td>
<td>1. Control (I) - 2. Win-Marking (J)</td>
<td>0.10</td>
<td>&lt;0.05</td>
<td>0.53</td>
</tr>
<tr>
<td>H1.2</td>
<td>1. Control (I) - 4. Challenge (J)</td>
<td>0.18</td>
<td>&lt;0.001</td>
<td>1.18</td>
</tr>
<tr>
<td>H1.3</td>
<td>1. Control (I) - 3. Juice (J)</td>
<td>0.09</td>
<td>&lt;0.05</td>
<td>0.53</td>
</tr>
<tr>
<td>H2</td>
<td>2. Win-Marking (I) - 6. Win-Marking vs. Challenge (J)</td>
<td>0.25</td>
<td>&lt;0.001</td>
<td>1.09</td>
</tr>
<tr>
<td>H3.1</td>
<td>2. Win-Marking (I) - 7. Win-Marking vs. Juice (J)</td>
<td>-0.07</td>
<td>0.37</td>
<td>0.27</td>
</tr>
<tr>
<td>H3.2</td>
<td>4. Challenge (I) - 8. Challenge vs. Juice (J)</td>
<td>-0.01</td>
<td>4.19</td>
<td>0.01</td>
</tr>
<tr>
<td>H3.3</td>
<td>5. Win-Marking + Challenge (I) - 9. Both vs. Juice (J)</td>
<td>-0.09</td>
<td>&lt;0.01</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Table 5.3: Planned comparisons to test hypotheses. The measure is mean relative preference calculated based on relative velocity as defined in the above section Measure, table 5.1.

This experiment tested the capacity for different play appeal factors to conflict, and thus pull players away from one activity that was arbitrarily designated the “correct” activity for the player to pursue (as designers of serious games would do when trying to develop a specific skill in players).
A repeated-measures one way ANOVA\(^6\) showed that there was a statistically significant difference between all conditions, \(F(8, 585) = 6.64\) (\(p < .001\)), justifying follow-up comparisons between specific conditions to test the hypotheses. The planned contrasts were made with two-tailed paired t tests with Bonferroni correction.

Introducing win-marking attracted players to pursue the indicated state, supporting the game value adoption concept of game motivation, \(t(66) = 3.04\) (\(p < .05\)). Cohen’s (1988) effect size value (\(d = 0.53\)) suggested a moderate practical significance. This supported H1.1 that players would pursue a win-marked game state over an unmarked one, derived from the concept of game value adoption implied in the game design literature.

As expected, introducing a difference in challenge resulted in players preferring the state that was more difficult to achieve, \(t(66) = 8.27\) (\(p < .001\)). Cohen’s effect size (\(d = 1.18\)) suggested a high practical significance. This supported H1.2 that players would pursue a challenging game state over a non-challenging game state, derived from the concept of challenge in game design literature.

Introducing juiciness succeeded in directing the player towards the intended win state, \(t(66) = 3.32\) (\(p < .05\)). Cohen’s effect size (\(d = 0.53\)) again suggested moderate practical significance. This supported H1.3 that players would pursue a juicy game state, derived from the concept of juice in the game design literature.

While all of the lose-tug effects were in the direction predicted, they were smaller than expected. One comparison where players clearly preferred the failure agent over the win agent was when challenge was introduced to lure players away from the win state.

\(^6\)Analysis of Variance (ANOVA) is a statistical test to determine if there is a statistically significant difference between more than two groups, before testing if a specific group has a statistically significant difference from another specific group. This is useful to reduce the rate of false positives. Testing for differences between two specific groups after an ANOVA is a \textit{post-hoc test}, and is usually done with some variation of t test. A full understanding of these statistical tests is not necessary to understand the results of these experiments and their implications (for more details on statistical techniques, refer to educational resources and textbooks such as Seltman, 2012). ANOVA produces an F statistic based on the data, the degrees of freedom in the data, the number of groups, and the sample size. The F statistic is then used to derive a p value, which is considered statistically significant if it is below .05. Such a result would indicate that post-hoc tests would be appropriate (if the ANOVA is not significant, then post-hoc tests are generally considered inappropriate). A t test to check for differences between specific groups will produce a t statistic based on the data and the sample size (similar to the F statistic of an ANOVA), which is then used to derive a p value. If that p value is less than .05 then the difference between the groups is considered statistically significant.
from win-marking, and this lose-tug effect was significant, \( t(66) = -8.68 \) (\( p < .001 \)). Cohen’s effect size value (\( d = 1.09 \)) suggested a high practical significance. This supported H2 that players would pursue a challenging game state over a state visually marked as positive, derived from the primacy given to the concept of challenge (Sherry and Lucas, 2004).

However, juice failed to tug significantly from win-marking, \( t(66) = -1.97 \) (\( p = .37 \); Cohen’s \( d = 0.27 \) suggested low practical significance), or to tug from challenge, H3.2, \( t(66) = -0.53 \) (\( p = 4.19 \); Cohen’s \( d = 0.01 \) suggested low practical significance). This results in the rejection of H3.1 that players would pursue a game state that is juicy over one that is challenging, and rejection of H3.2 that players would pursue a game state that is juicy over one that is visually marked as positive, derived from the conjecture that juice may overpower these other play appeal factors (Malone, 1980; Ravaja et al., 2005).

Interestingly the lose-tug effects of juice on the combination of challenge and win-marking was significant, \( t(66) = -3.56 \) (\( p < .01 \); Cohen’s \( d = 0.5 \) suggested moderate practical significance). This result supports H3.3 that players would pursue a game state that is juicy over one that is both challenging and visually marked as positive, derived again from the conjecture that juice may overpower these other play appeal factors. How this result should be interpreted is not clear, given that juice’s effect on each of those in isolation was not strong. It suggests that the interactions between these motivations might be non-linear, which would conflict with the hierarchical theory of play-motivation-interaction. This would imply that this phenomenon of ‘why do players choose to lose’ is more complex than originally supposed. These results are summarised simply in table 5.4. Details can be found in table 5.3, with a visual indication of the size of differences between conditions in figure 5.10.
## Results

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Description</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>Challenge, juice, and win-marking are each, in isolation, capable of attracting players to pursue them</td>
<td>Supported</td>
</tr>
<tr>
<td>H1.1</td>
<td>Players will pursue a win-marked game state over an otherwise identical unmarked game state (game value adoption)</td>
<td>Supported</td>
</tr>
<tr>
<td>H1.2</td>
<td>Players will pursue a challenging game state over an otherwise identical game state that is not challenging</td>
<td>Supported</td>
</tr>
<tr>
<td>H1.3</td>
<td>Players will pursue a game state with gratuitous feedback over an otherwise identical game state with basic feedback (juice)</td>
<td>Supported</td>
</tr>
<tr>
<td>H2</td>
<td>Challenge Beats Win-Marking: Given a choice between two states, players will pursue the state they find more challenging even if the other state is visually marked as the win state</td>
<td>Supported</td>
</tr>
<tr>
<td>H3</td>
<td>Juice Reigns Supreme: Juice takes priority over challenge and win-marking, capable of attracting players away from both.</td>
<td>Rejected</td>
</tr>
<tr>
<td>H3.1</td>
<td>Players will pursue a juicy game state over a win-marked game state</td>
<td>Rejected</td>
</tr>
<tr>
<td>H3.2</td>
<td>Players will pursue a juicy game state over a challenging game state</td>
<td>Rejected</td>
</tr>
<tr>
<td>H3.3</td>
<td>Players will pursue a juicy game state over a challenging win-marked game state</td>
<td>Supported</td>
</tr>
<tr>
<td>Hierarchical Play Interaction</td>
<td>Play motivators can be arranged in a hierarchy with ones above being pursued over the ones below. Juice sits above challenge, which sits above win-marking</td>
<td>Rejected</td>
</tr>
</tbody>
</table>

Table 5.4: Summary of the results of this study. Details can be found in the above tables and figures.

### 5.6 Discussion

This study investigated RQ2: How can different combinations of game features interact to cause players to engage in off-task behaviour instead of trying to win the game (and thereby likely miss the purpose of an educational or serious game)? This was inspired by various observations (discussed in the literature review above) of instances where players pursue activities irrelevant to the goal of the game, or even seem to enjoy failure. This prompts one to consider, why might players choose to lose?
This is an important family of phenomena to study considering how problematic it can be when players pursue a game state that is not the intended goal of the game, and there are many ways it could be approached. For example, one could assume that this is deliberate transgressive behaviour. But this study sought to determine if a cause in the game design (rather than intent in the player) could provide a partial explanation toward some of these phenomena. That is, even if a player is not specifically (knowingly) trying to fail the game, could a flaw in the game’s design cause them to pursue activities other than the goal? This study refined the broad question, ‘why do players choose to lose’, into a more precise form, ‘can a difference in a game design result in conflicting play motivations that pull players away from the win state? If so, which play motivations win out when they conflict?’

This study investigated this phenomenon by putting prominent game design concepts to the test and testing a hierarchical theory of their interaction via two main hypotheses. This required the development of a model that could quantify the challenge of different states and generate predictions of player behaviour. The Dynamic Probability Response model proved effective in quantifying the challenge of the different conditions and also in generating accurate predictions about player behaviour based on those measurements.

The foundational hypothesis (H1) was that all three of these concepts were indeed valid. The results supported all three concepts of play motivation from the game design literature. Players in this study tended to pursue game states that are difficult to achieve, visually indicated as win states, and states that provide gratuitous feedback. These findings support the challenge, game value adoption, and juice concepts, respectively. The latter two of which had been with little to no direct experimental support until now (Although, see Hicks et al., 2019; Juul and Begy, 2016).

The results concerning why players might choose to lose are more complex. Lose-tug effects were observed and statistically significant, indicating it is possible for a game designer to accidentally make one state more appealing than their intended win state, indicating this approach (investigating game design flaws...
as causal factors in off-task behaviour) has validity. This indicates that, regarding RQ2, different combinations of game features can indeed interact to cause off-task behaviour. Specific combinations did interact to produce this effect, but these results also highlighted the complexity of those interactions (for a summary, see table 5.4; for details see table 5.3; for a visual indication of the size of differences see figure 5.10), requiring more research to fully understand.

The literature review above implied a simple hierarchical interaction, where play motivators higher in the structure overpowered those below. Specifically, the observations of Malone (1980) and Ravaja et al. (2005) implied that juice overpowers both challenge and the visually indicated win state of the game, putting it above both of those in the hierarchy (H3). Challenge was hypothesised to be next down in the structure, above win-marking but below juice (H2). This was because the above authors explicitly placed juice above challenge, but aside from this, challenge is often assigned high relative importance in play motivation literature (see e.g. Sherry and Lucas, 2004). This left win-marking at the bottom of the hierarchy. However, this simple hierarchical model of play interactions was not supported, meaning more work is needed to develop theories to explain and predict play motivation interactions.

There was support for the hypothesis that challenge can overpower the indicated win state (game value adoption) under certain circumstances (H2). However, the interactions of juice with the other play motivators was not consistent with any place in the hierarchy. This study found support for the hypothesis that juice could pull players away from the combination of challenge and win-marking, however juice did not have a significant effect on either challenge or win-marking when they were not combined. Therefore, this study was unable to find support for the hypothesis of Malone (1980) and Ravaja et al. (2005) that juice sits at the top of the hierarchy, and casting doubt on the notion of a hierarchy of interaction as there is no place in a hierarchy where this combination of results would be predicted. This suggests play motivators can interact in complex, non-linear ways to amplify or dampen each other, and much additional work is needed to develop a theory of such interactions. While each of the concepts of game motivation (challenge, juice, and game
value adoption) were supported in isolation, this field does not yet have any good theory to predict how these motivations interact.

Regarding the more practical concerns of RQ2, the above results supported the contention that players may prefer to lose the game if the design is flawed and player motivations conflict, and suggest two possible answers to why players choose to lose:

1. The lose state may be more challenging even when the win state is marked.

2. The lose state may be juicier even when the win state is both more challenging and marked as the win state.

These answers to RQ2 have practical implications for serious game design, which are discussed below. Once again, note that this study did not demonstrate that this is a universal law, but that it can happen under the right circumstances.

It is important when interpreting these results to be mindful of the limitations of this study as one of the few directly investigating competition between game play motivators. More studies looking at other combinations of motivators in different contexts are needed before any definitive and widely generalisable conclusions can be drawn. This study used one kind of game, and represented only one way to achieve each motivator: There are many other game genres, and many other ways to visually indicate the win state, or to provide juice, or to create challenge. There are also likely to be interesting interactions between these lose-tug effects and individual traits such as demographics, approach or avoidance motivational style, and play style preferences, but these are beyond the scope of this study.

Therefore, it is important to remember that this study didn’t demonstrate that challenge always takes priority over the visually-marked win state in all types of games with all types of players (nor did it set out to do so), but merely demonstrates that it can happen (specifically, under circumstances similar to those seen in this study). This doesn’t provide a final and universal answer to why players choose to lose but the results do indicate game designers should be vigilant to avoid these competing motivators defeating the purpose of their game, and demonstrates an
experimental design and methodology that can be used in future studies to test the generalisability of these findings and to explore related questions.

5.6.1 Play Motivations Application & Theory

The findings suggest future directions for research and have implications for the design of serious games and simulations.

If future research can replicate this effect, it suggests care must be taken to ensure the communicative cues in the game correctly convey to the player what they should be pursuing. Designers need to consider not only the functionality of their games and simulations, but also how the systems are represented and their connotations. In this study players preferred to pursue a positively-represented game token even after the acclimatising sessions demonstrated that it does nothing differently from the neutrally-represented token. It seems players might be influenced by the associations of representations even after learning the game mechanics. This is an unexpected result considering previous findings that people learn to ignore extraneous details (see Haider and Frensch, 1996; Juul, 2005; Rambusch and Susi, 2008; however, see also Bateman, 2013a; Bateman, 2013b). But it gives support to the game value adoption concept of game motivation which predicts players will follow cues from the game about which are the ‘correct’ states to pursue. However, questions still remain about how best to communicate the goal and to avoid miscommunication, and about various other components and predictions of game value adoption, as discussed in the literature review section Game Design.

If future research is able to replicate and extend the results regarding juice, then it suggests designers need to consider the magnitude and spectacle of the reactions the system provides in response to player’s successful and unsuccessful choices. It may be realistic for emergency services to scramble to the scene and crowds to scatter in a screaming panic, but such an interesting reaction might reward players and pull them away from successfully managing the situation. It may be worthwhile to omit or subdue a realistic response to failure and exaggerate the response to success so that it provides greater spectacle.
The importance of challenge as a psychological motivator is well-established. But game designers and researchers lacked a way to quantify challenge (both positively and negatively valorised) independent of the mechanics used to generate it. Dynamic Probability Response proved effective for both measurement and prediction. For designers, it is a tool that can be incorporated into the growing effort to record player metrics in game development to make game design more data-driven and less subjective. For researchers, it can allow studies to be compared even with different contexts for generating challenge. Furthermore, as other concepts of play motivation are developed to the point of quantified measurement, it will also allow researchers to find algorithms that describe interactions between motivators such as disruption or amplification. The results supporting the juice and game value adoption concepts demonstrate that it is worthwhile for future studies to develop these concepts further until they enable similar objective measurement and comparisons.

Future work should also concentrate on investigating the various other play motivations in the game design literature. There are many proposed taxonomies of play and fun (for reviews see Bateman, 2014; Hamari and Tuunanen, 2014), but few theories of how they might interact. The most common paradigm (implicit in the explanations in e.g. Malone [1980] and Ravaja et al. [2005]) is a simple hierarchy where one overpowers those below it. This study found support for the claim that challenge can overcome game value adoption under the right conditions, but was unable to find support for juice’s hypothesised place at the top of the hierarchy as was supposed. As we develop empirical support and theoretical models for individual motivators in games, we need also to develop empirical data and theoretical models of how they interact with each other: which ones can cancel each other out, amplify each other, or not affect each other at all? There is currently a dearth of empirical data on which to base theories of play-motivation-interaction, making it a good starting place for future work.
5.6.2 Limitations & Future Work

Caution should be taken generalising these findings. This study involved very brief play times and there might be different results with longer-term play behaviour. For instance, Bartle (1996) observed differing player preferences for activities and an apparent shift in those preferences over time. But the shift which he discussed took place over a matter of months or more, and the duration of play in this experiment was a matter of minutes, making it unlikely that the kinds of changes needed for these shifts would have had time to occur in this study. Future longitudinal studies could be designed to detect a shift in player interest to see if there is a pattern in these shifts consistently towards certain kinds of gaming activities that are suboptimal or harmful to achieving the win state. Such individual differences as personal play preferences and other demographic details such as background experience with games could play a large part in determining how different play motivations interact. Therefore the absence of such measures in this study is another limitation and promising direction for future research.

Additionally, only one game type was used, and the variety of game genres is vast. It could be the effects observed in this study are specific to the chase genre. Theories of challenge assume that what matters are not the specific mechanics but the degree and type of challenge they generate. Nevertheless, future studies need to investigate if similar effects can be found in different game genres.

Similarly, win-marking was a heart icon, and the results may have differed had other symbolism been used. Many games use coins, cherries, diamonds, or other more complex symbols to convey this, such as damsel in distress symbolism, or a fully-fledged narrative with backstory in order to convey which condition is the win condition. Many such win-marking symbols have very different associations. For example, a heart icon often represents health or extra lives, whereas coins often represent currency for purchasing items and upgrades. Although the assumptions of the game value adoption concept define these methods as equivalent, empirical studies should test this assumption. The results of this experiment suggest further development of the game value adoption concept may be worthwhile. A salient
STUDY 1: EVALUATING PLAY APPEAL FACTORS & ENGAGEMENT

place to start is to determine what are the minimum necessary conditions for players to adopt the values of a game.

Juice was achieved by using one combination of effects every time (with random variation). Future research could explore different techniques to create juiciness. Although the assumptions of the juice concept are that any method of achieving a spectacular reaction is equivalent, more empirical work is needed to determine if this is so. Future work on creating a model of juice is necessary to allow us to quantify just how juicy something is, and to predict how this will affect player behaviour.

This study approached the question, ‘why do players choose to lose,’ from one of many possible angles. Future studies should investigate other possible approaches to this question, such as those that involve deliberate transgression (e.g. spoilsporting and comedic transgressive play). Here, I examined a best-case scenario to see if players could be trying to win, but misinterpret the goals of the game (and to see which motivations win out when such conflict exists). However, this does not mean that this is the only or the primary factor to explain choosing to lose as it naturally occurs in the wild. Other possible causes should be investigated in future work.

I have identified three possible play motivations to study. Each of them may be present or absent for achieving a particular game state, creating an array of possible combinations – a game state may be challenging, but not juicy and with no win marking, or any combination. There are also two game states (e.g. the win state and the lose state) to consider. Therefore, another array of possible combinations of motivators exists for the second game state in the study. Any one of the possible combinations for the first game state could be compared to any possible combination for the second game state, creating another layer of possible combinations. Furthermore, the three motivators examined herein are not exhaustive, and many other possible play motivators could be included. A study could even include a third or fourth game state, further increasing the combinatorial possibilities. Therefore, it is worth noting that this study only makes use of small selection of all possible motiva-
tor combinations for game states in order to test a specific set of hypotheses. Future studies could test alternate theories of play motivation interaction that require a different selection of motivator combinations from the vast set of possibilities.

5.7 Integrating With MMT

This experiment tested whether different play appeal factors \((a_g)\) could have a non-linear interaction such that play appeal factors present in the same game could interfere with each other rather than reinforce and augment each other in formula 3.15. This idea was supported, which highlights the importance of the fact that this field currently lacks any established theory of how different play motivations or play appeal factors may interact with each other. Such a theory would be able to predict engagement based on a function that included multiple different play appeal factors, instead of just one \((f(a_g))\).

DPR joins just a small handful of other theories (reviewed above) of play appeal factors that have been formalised to the point of an operational definition. These few enable the objective quantification of a play appeal factor (such as challenge or agency) in a game. The measured play appeal factor can then be used in MMT formula 3.15 to derive predictions of engagement. DPR doesn’t specify a detailed relationship between the play appeal factor of difficulty-to-achieve and engagement \((f(a_g))\), simply predicting that an increase in the factor will result in an increase in engagement. For theories that do not specify a complex function relating the factor to engagement, it would be reasonable to proceed on the assumption that it is linear. In contrast, the theory of flow does the opposite job: It predicts a specific U-shaped relationship between difficulty and engagement, but it fails to specify objectively quantifiable play appeal factors that would correspond to difficulty. Thus, different theories can be brought together in MMT to augment and extend each other. For example, DPR could provide an operational definition of difficulty as a play appeal factor \((a_g)\), and flow could be used to provide the function specifying the relationship between that play appeal factor and engagement \((f(a_g))\).
This line of investigation is pursued in the next experiment, where complexity theory is used to provide operational definitions of play appeal factors, and flow provides the predicted relationship to engagement.

5.8 Conclusion

This study investigated RQ2: How can different combinations of game features interact to cause players to engage in off-task behaviour instead of trying to win the game (and thereby likely miss the purpose of an educational or serious game)? In so doing, it contributed to the field empirical findings of how certain combinations of game features (juice, challenge, and win-marking) can interact with each other to cause such problems. Specifically, the findings suggest two possible answers to why players choose to lose:

1. The lose state may be more challenging even when the win state is marked.
2. The lose state may be juicier even when the win state is both more challenging and marked as the win state.

Many individual motivators have been discussed in the game design literature, but few have been empirically tested. It appears that the intuitions of game designers can provide psychological insights. One of the aims of this study was to assess the validity of common game design concepts like juice and game value adoption in game design literature. These preliminary findings support the concepts according to their basic definitions provided above. This support suggests they deserve further development to take them from basic definitions to formal scientific models. Specifically, juice could be further expanded and studied in terms of dimensions such as salience, complexity, malleability, and permanence of effects. Game value adoption could be elaborated in terms of exactly what cues are necessary to convince people that a context is a game, and how exactly this renders people willing to adopt arbitrary goals or values.
Models to quantify juice, game value adoption, and other gameplay motivations could prove invaluable considering the Dynamic Probability Response model of challenge developed for this study proved effective in quantifying the challenges present in the experiment and making accurate predictions. The DPR model of challenge contributes a method to objectively quantify the degree of challenge in different conditions, and therefore can facilitate establishing connections between specific game properties (e.g. challenge) and outcomes such as engagement and learning. It will be necessary to conduct further studies to test the assumptions and predictions of Dynamic Probability Response (e.g. that players will report feeling diminished agency the flatter the challenge curve is).

If other studies can independently replicate these results with different game mechanics and genres and test other predictions generated by Dynamic Probability Response, it could become a useful tool for game designers. The recent growth of data-driven design shows that scientific approaches are sought after in game and simulation design. Serious games such as those developed for education and rehabilitation need to be underpinned by scientifically derived patterns and models of player behaviour, experience and learning in order to understand how players will respond to the structure, dynamics and aesthetic representations of games.

This experiment demonstrated how the MMT framework can guide experimental design to probe more deeply and precisely into specific phenomena and questions in educational and serious game design, while still retaining a place in MMT and therefore relevance to the broader picture of how educational and serious games can achieve positive outcomes. It also demonstrated how MMT can be expanded and elaborated with more detailed models of specific components - in this case, the DPR model of challenge to specify the relevant game properties to measure (forming the play appeal factors in MMT’s equations), and generate predictions of engagement.

These findings on engagement can be directly inserted into MMT to elaborate the process of Mastery. Improved engagement should result in improved mastery of the game and, assuming high veracity of the game, consequently improved
transfer to the real world. Thus, the next logical step was to investigate if engagement and mastery co-occur under the right conditions, and investigate the Inherently Learnable Systems hypothesis.
Chapter 6

Study 2: Game Complexity & Curiosity

The second question to which MMT was applied was RQ3: What is the relationship between the complexity of a game, and the player’s ability to master it? What form of complexity best arouses the curiosity of players? This study explores the interaction of curiosity and complexity, and will also help to demonstrate the value and the manner in which MMT can be used to design an empirical experiment to probe a specific psychological phenomenon in educational games research, contributing empirical findings as to how measurable forms of game complexity relate to curiosity.

This study investigated whether measures of game complexity adapted from the complexity theory literature could function in MMT’s formulas 3.15 ($\tilde{e}_p = f(a_g)$) and 3.18 ($\tilde{\rho}_p = l(e_p, c)$) as both a play appeal factor ($a_g$) and a learning difficulty variable ($c$) for predicting outcomes of engagement and learning, and whether curiosity had a place as a play appeal preference variable ($a_p$) in formula 3.16 ($\tilde{e}_p = f(a_g, a_p)$). The results supported the relevance of these variables, but showed that more work is needed on the functions relating them to the outcomes ($f(a_g, a_p)$ and $l(e_p, c)$).
6.1 Introduction

This study implements guidelines from MMT to design the research approach taken to address RQ3. This study narrows the focus to take a deeper look at just one part of the process: How players develop mastery of a game, and specifically how games elicit interest by provoking curiosity. The aim of this approach is not to test a specific educational game to teach a specific subject, but to test an hypothesised mechanism or process to derive principles based on objective measures of game properties that will hopefully generalise across domains exhibiting similar properties. To ensure one’s findings are not just an artefact of using a specific educational subject, or the result of cultural associations and background knowledge, the focus is shifted to the abstract properties of games as systems. Semantic embedding (when a simulation depicts a recognisable real-world system) can hinder learning by fostering a set of presumptions that may not be accurate and tend to go untested (Beckmann and Goode, 2014). Goode (2011) therefore advises creating simulations that are abstract, so that they do not have a recognisable, familiar real-world counterpart. By eliminating obfuscating variables related to the narrative and world of the game, and focussing on analysing quantifiable variables of the game as a dynamic mathematical system, one can ensure that the results concern game systems generally and are not just due to cultural associations and background knowledge of a particular domain.

This study investigated the following research questions:

RQ3.1: What is the relationship between game system complexity and engagement?

RQ3.2: How is that relationship affected by individuals’ trait curiosity?

From the literature on game design, learning, and complexity theory, two broad claims were derived and tested:

1. Inherently Learnable Systems: An inherently fun system (i.e. a “game”), is an inherently learnable system (able to be self-taught without external guidance).
Therefore, conditions which result in high engagement, will also result in high learning.

2. Epistemic Emergence: A medium degree of complexity gives rise to emergence in the form of ‘simple rules causing complex behaviour’. The apparent conflict between rules (static structure) and behaviour (dynamics) is likely to cause separation of implicit and explicit learning – where implicitly mastering the behaviour of the system does not improve one’s ability to map or articulate the underlying rules. Therefore, a non-emergent condition will result in similar levels of implicit and explicit learning, whereas an emergent condition will result in different levels of implicit and explicit learning.

Also derived from the literature were several competing hypotheses regarding play, curiosity, and complexity (each based on different theories - Whether the results support or conflict with each of the below will help determine which of these theories is likely to prove effective in modelling the relationship between game complexity and engagement):

1. Emergent Gameplay: Simple systems and overly complex systems are similarly uninteresting, and the most interesting systems are those with a medium level of complexity, giving rise to emergent gameplay. This is based on the theory of play as flow, and theories of curiosity as incongruity or optimal arousal.

   (a) Edge of Chaos: Dynamic complexity (e.g. sensitivity to initial conditions) is its most interesting in between order and chaos, at the hypothesised ‘edge of chaos’.

   (b) Scale-Free Network: Static complexity (e.g. the causal structure of the game’s rules) is most interesting not when the amount of causal linkages is not minimal, and not maximal, but an intermediate amount that forms a scale-free topology (as seen in the ‘small world’ networks often considered complex when found in nature).

2. Challenging Complexity: High levels of complexity will be the most engaging, based on play as challenge, and the drive theory of curiosity.
3. Effectance: Low levels of complexity will be the most engaging, based on play as effectance.

4. Curious About Complexity: Drive theories of curiosity imply that individuals with higher curiosity as a character trait will be more interested in higher levels of complexity.

This study first derives the above competing hypotheses to test from examining the literature on educational games, play, curiosity, and complexity with respect to the research questions. Then, the experiment’s methods, procedure and results are explained. Finally, a discussion of the findings’ theoretical implications and practical applications leads into the conclusion. This study found no support for the flow theory of play (although this was not a direct test of flow, as it is a large theory with many elements other than challenge/complexity level), or the incongruity, optimal arousal, or drive theories of curiosity, but did find support for the effectance theory of play. The experiment also explored the role of curiosity as a character trait in determining how interesting individuals found different complexity levels, testing the drive theory of curiosity which would predict that more curious individuals would be drawn to more complex systems. This hypothesis was not supported, the data instead indicating that more curious individuals are generally more interested in all complexity levels than less curious individuals, and suggesting that educational game design should focus on catering towards the less curious who could be more sensitive to differences in complexity level.

This is an important initial step in discovering which theories of play and curiosity are relevant to complex games, and which conceptions of complexity are appropriate and useful for understanding engagement and learning in complex games. The theoretical implications, practical applications for educational game design, and limitations and future directions for research are discussed.

To distinguish the experiments of this second study from the first study in the previous chapter (which can be regarded as Experiment 1), these two new experiments will be referred to as Experiment 2.1, and Experiment 2.2, to demarcate that they belong within Study 2.
6.2 Hypotheses From Theories

The literature review section of this thesis described various theories which would make certain predictions concerning the outcomes of this study.

6.2.1 Emergent Gameplay & Epistemic Emergence

Emergent gameplay is frequently discussed as a core design feature in games that promote exploratory play (Kickmeier-rust and Albert, 2009; Sweetser, 2008). A common definition of emergence is a simple set of rules that give rise to complex behaviour (Crutchfield, 2011; Jost, Bertschinger, and Olbrich, 2010; Garneau, 2008a; Sweetser, 2008). Unexpected phenomena pique one’s curiosity (Berlyne, 1954; Berlyne, 1966; Day, 1982; Loewenstein, 1994), which is a central motivation of exploratory play, making this a promising concept for designing inherently learnable systems. Combined with Bonawitz et al.’s (2011) suggestion that exploration is better suited to contexts where the teacher lacks the time or ability to provide comprehensive instructions, this suggests a complimentary relationship between these concepts: Emergence is well suited to elicit exploratory play, and exploratory play is well suited to learn about emergence.

This definition of emergence implies that the behaviour of an emergent system can be surprising, even when one has full knowledge of the rules by which it operates. This implication can be combined with dual process theories of cognition (Evans, 2012; Kahneman and Klein, 2009; Osman, 2004; Smith and DeCoster, 2000; Sun and Zhang, 2004), which posit that implicit learning can occur largely independently of explicit learning. The combination of dual process theory with the above definition of emergence produces the prediction that an emergent system will result in implicit learning of its behaviour occurring at a rate that is independent of explicit learning, and vice versa. Therefore, if a change to a system results in implicit and explicit learning uniting and co-occurring, then that change can be considered to have removed a feature that is necessary for emergence. Conversely, if a change to a non-emergent system (where implicit and explicit learning co-occurred) results
in a separation of rates of implicit and explicit learning, then that change can be considered to be sufficient for emergence. According to this premise, the separation of implicit and explicit learning can be used to identify system properties that are sufficient for emergence, and the uniting of implicit and explicit learning can be used to identify system properties that are necessary for emergence.

This research shall remain agnostic as to the actual mechanism behind implicit and explicit reasoning (see literature review chapter above, section *Implicit vs. Explicit Theories of Learning for Games*). Whatever they may be, the important thing for this research is that this definition of emergence implies that explicit learning of rules and implicit mastery of behaviours should only separate into independent learning processes when there is sufficient complexity – and that therefore explicit and implicit learning can be reunited into the one process at lower levels of complexity. This as yet untested prediction of the separation of implicit and explicit learning in emergent systems I will term the *Epistemic Emergence Hypothesis*: Implicit and explicit knowledge scores will be different (e.g. high implicit but low explicit) in the emergent complexity condition. This outcome would be consistent with the definition of emergence as simple rules giving rise to complex behaviour (a difference between static and dynamic levels of complexity).

There are many definitions of complexity, such as the balance between change and stability (Fernandez, Maldonado, and Gershenson, 2014), self-dissimilarity at different scales (Wolpert and Macready, 2004; Wolpert and Macready, 2000), among others (for reviews, see Bonchev and Buck, 2005; Chu, Strand, and Fjelland, 2003). With no consensus on which one is “correct”, it is left to the discretion of individual researchers to select (or develop) a complexity measure that is most suited to answering their research question. This study requires two definitions of complexity in order to distinguish the “simple rules”, from the “complex behaviour” – which I will call *static complexity* and *dynamic complexity* respectively. *Dynamic complexity* is any form of complexity that is conceptualised or measured in terms of how the system’s state changes over time (i.e., the system’s behaviour, patterns of output, or how it responds to different starting conditions or interventions), requiring simulation or observation of the system in operation. This is to distinguish it from *static*
complexity, which is any form of complexity that is conceptualised or measured in terms of fundamental elements, structure, or descriptions of the system that remain unchanged as the system is active (e.g. the number of components, the information represented by the system’s rules, the structure of its data, etc.) and therefore do not require tracking system state changes over time.

Dynamic complexity requires a measure focusing on the behaviour of the system. Emergence has been described as occurring at the edge of chaos (Adams and Dormans, 2012; Brodu, 2008; Chalmers, 2002; Chi et al., 2012; Fromm, 2006; Hudson, 2011; Prokopenko, 2013; Sweetser, 2008), putting emergence in between order and total chaos (figure 2.4). The presence of certain properties like topological mixing and strange attractors are signs of chaos (Mitchell, 2009; Sterman, 2000), but the central element to chaos is sensitivity to initial conditions (Gleick, 1988; Nunn, 2007). This can be measured along a continuous scale using the Lyapunov exponent (Cross, 2000; Habib and Ryne, 1995; Rosenstein, Collins, and De Luca, 1993), which will be this study’s measure of dynamic complexity (see literature review chapter above for more details, page 66. But a detailed understanding is not necessary to understand this study).

Static complexity of a game system (the complexity of its rules) can be measured with Kolmogorov complexity, as explained by Tornqvist, Wen and Tichon (2017a). However, there is another measure of static complexity that converges with others in predicting that emergence should arise in the middle ground between the extremes of complexity.

Network theory contains much discussion of static forms of complexity. Since game systems are often represented as a network of objects connected by lines of interaction (Adams and Dormans, 2012; Smith and Smith, 2004; Zupke, 2016), network theory is an apt domain from which to draw a measure of the complexity of game rules. Similar to the edge of chaos hypothesis, network complexity is considered highest at an intermediate level of connectivity: When there is minimal connectivity, the network is considered relatively simple, but if there is maximal connectivity (i.e. every node is connected to every other node) then the network is
also considered quite simple (Barabási, 2002; Bonchev and Buck, 2005). Some level of connectivity between minimal and complete connectivity is required to create a complex network (figure 2.5).

Complex naturally-evolving networks (e.g. the internet) tend to be scale-free networks (Barabási, 2002; Mitchell, 2009; Wen, Kirk, and Dromey, 2007; Wen, Dromey, and Kirk, 2009). Since the structure of scale-free networks is considered complex in network theory, it is worthwhile to investigate if games with such structures tend to result in emergent gameplay.

Total walk count (TWC) is a measure of connectivity that involves counting the number of unique paths that can be traced between each pair of nodes (see related literature review above for more information, page 68. Detailed formulas can be found in Bonchev and Buck, 2005, pp. 211-213. But a detailed understanding is not necessary to understand this study). This measure could be used to determine if an intermediate level of connectivity is the most engaging to learn. Note that, as discussed in the literature review, high connectivity would indicate high static complexity according to a naive conception of network complexity, whereas one based on information theory would consider complete connectivity to be low static complexity. Therefore, since the naive perspective on connectivity produces the inverted U-shape concordant with the other theories used in this study, it will be used as the default perspective in this study (and contrasted later in the discussion with other perspectives that consider complete connectivity as simple instead of complex).

Both the “edge of chaos hypothesis” of dynamic complexity and the “scale-free network” hypothesis from network theory bear a striking resemblance to a concept often discussed in the game design literature on how challenge can motivate play: Flow (Adams, 2009; Aponte, Levieux, and Natkin, 2011; Brockmyer et al., 2009; Klarkowski et al., 2015; Koo, 2009). There are many components to this theory, but one of its central themes is the prediction that flow occurs at an intermediate level of challenge: Too little difficulty results in boredom, but too much results in frustration. All three of these theories from different fields – Flow, edge of chaos, and scale-free networks – have independently converged on very similar predictions of
an inverted U-shaped graph of inherently learnable systems (figure 6.1). I refer to this claim, that an intermediate degree of complexity (such as the edge of chaos or a scale-free topology) will result in a more interesting and engaging, emergent system as the *Emergent Gameplay Hypothesis*: Higher engagement will occur in the medium complexity condition, and lower engagement in the low and high complexity conditions. This outcome would be consistent with the theories of play as flow and curiosity as optimal arousal, and consistent with the incongruity and knowledge gap theories of curiosity on the condition that the edge of chaos and scale-free network theories of emergence hold true.
Figure 6.1: The Edge of Chaos, Scale-Free Network, and Flow theories all suggest an inverted U-shaped relationship between some form of complexity on the x axis, and some other variable on the y axis.
6.2.2 Play & Curiosity

Various theories of play and curiosity were described in the literature review chapter of this thesis, Game Design section. The theories reviewed provide conflicting predictions for the outcomes of this study, concerning how complexity affects engagement. The relevant theories drawn on are play as challenge, play as effectance, and the following theories of curiosity: drive theory, optimal arousal, knowledge gap and incongruity.

Players motivated by challenge might relish high levels of complexity and seek to master them. I term this the Challenging Complexity Hypothesis: Higher engagement will occur in the higher complexity condition, and lower engagement with lower complexity. This would be consistent with the theory of play as challenge.

The literature on juice and effectance produce different predictions for this study. Play just for the immediate gratification of seeing the game react to one’s actions would be less concerned with mastering high levels of complexity. A less complex system may empower and enable such immediate gratification by having simple and easy ways to have obvious and intuitive effects on the game by direct interaction. I term this the Effectance Hypothesis: Higher engagement will occur in the low complexity condition, and lower engagement in the high complexity condition. This outcome would be consistent with the theory of play as effectance.

The form of play most pertinent to this study is exploratory play (Tornqvist, 2014), which is characterised by spontaneous interventions to resolve ambiguity.

Curiosity is considered one of the primary motivations of exploratory play (Keller, Schneider, and Henderson, 1994; Loewenstein, 1994; Schulz, Bonawitz, and Griffiths, 2007; White, 1959; Lorenz, 1981). It is considered both a trait and a state (Grossnickle, 2016), meaning that curiosity is both a stable characteristic that can vary across individuals, and a state of mind that can vary over time within an individual. As a character trait, curiosity could be a central causal variable of the player in the process of playing and mastering complex games, as more curious individu-
als may be more intrigued by complexity in games, or more disposed to exploratory play.

Drive theories suggest that those with higher trait curiosity have a stronger need for harder intellectual problems to solve, and therefore would be more engaged with higher complexity conditions. This is identical to the Challenging Complexity Hypothesis defined above: The higher the complexity, the higher the engagement. Such an outcome would be consistent with both the theory of play as challenge, and curiosity as a drive. But it also produces a more specific hypothesis I will term Curious About Complexity Hypothesis: Individuals with higher trait curiosity will be more engaged with higher complexity conditions.

Optimal arousal theories imply that medium levels of complexity form the optimal range of stimulation and would therefore be preferred by players. This produces the hypothesis that higher engagement will occur with a medium level of complexity, and lower engagement with low complexity and high complexity – This is an hypothesis consistent with ideas on Emergent Gameplay, and will therefore be called the Emergent Gameplay Hypothesis (see below section). Knowledge gap and incongruity theories would predict that a system with surprising, emergent behaviours could provoke curiosity.

6.2.3 Exploratory Hypotheses

Many of the above theories produce ambiguous or contradictory predictions for the outcomes of this study. Therefore, this study will proceed with broad, exploratory hypotheses that minimise assumptions about the underlying theory.

These various theories of play (challenge, effectance, flow, exploratory), curiosity (drive, optimal arousal, incongruity), and emergence (edge of chaos, scale-free network) all produce conflicting predictions about how the complexity of game systems relates to engagement, but they all agree that there should indeed be some kind of relationship. With no particular reason to favour one theory over another at
this stage of research in this field, this produces the exploratory, primary hypothesis of this study:

H1: Game system complexity influences engagement.

Discovering that relationship between game system complexity and engagement will illuminate which theories of play, curiosity, and emergence might be more appropriate to the domain of educational games and teaching about complex systems.

More specifically, the various theories of curiosity as a drive, optimal arousal, incongruity or knowledge gap, all produce different predictions of how curiosity as an individual character trait will affect engagement with complexity. Without reason to favour one theory of curiosity, this produces the exploratory secondary hypothesis of this study:

H2: Trait curiosity influences engagement with complexity.

Discovering how trait curiosity affects engagement with complexity will provide further information as to which theory of curiosity might be more appropriate for this domain of study.

Total walk count and Lyapunov exponent can therefore serve as measures of static and dynamic complexity, respectively. These two forms of complexity have an intuitive correspondence to the two forms of knowledge that should be required to understand them: Explicit and implicit knowledge. This is of particular interest for an emergent system, which is of such complexity that its full implications cannot be intuited even from complete knowledge of its structure.

The reviewed theories of cognition do not make specific use of a distinction between static and dynamic complexity, making the exact relationship between these types of complexity and learning unclear under dual-process theory: Will implicit learning be high and explicit learning low, or vice versa, and will it be the same or different for static versus dynamic complexity? Therefore, the third major hypothesis of this study is also exploratory:
H3: Game system complexity influences learning of that game system.

Discovering the nature of static and dynamic complexity’s relationships to implicit and explicit learning will be necessary to developing informed theories of how and why different types of complexity result in different kinds of learning.

It is at this intersection point, between order and chaos, between boredom and frustration, and between explicit and implicit learning, that one can find the most fertile ground to test these related hypotheses from different scientific domains (complexity theory, game design, and cognitive science) and in so doing, help inform and develop all three domains in one fell swoop.

6.3 Method: Experiment 2.1

This experiment was designed to investigate the two major hypotheses of this study, H1 (Game system complexity influences engagement) and H3 (Game system complexity influences learning of that game system), as defined and explained above.

A between-subjects design was used, with each participant randomly assigned to a different level of static or dynamic complexity, making for a total of six groups (table 6.1).

<table>
<thead>
<tr>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1: Low Dynamic Complexity</td>
</tr>
<tr>
<td>Group 2: Medium Dynamic Complexity</td>
</tr>
<tr>
<td>Group 3: High Dynamic Complexity</td>
</tr>
<tr>
<td>Group 4: Low Static Complexity</td>
</tr>
<tr>
<td>Group 5: Medium Static Complexity</td>
</tr>
<tr>
<td>Group 6: High Static Complexity</td>
</tr>
</tbody>
</table>

Table 6.1: Between-subjects design of experiment 2.1.

1It is also possible to create a three-by-three matrix of static and dynamic complexity, enumerating every combination of low, medium, or high dynamic complexity with low, medium or high static complexity, for a total of nine groups. However, nine groups would create a particularly large empirical experiment that would require a very large sample size to achieve statistical significance. Secondly, it has not yet been established how dynamic and static complexity may interact (they may not be orthogonal dimensions). Thirdly, it has not yet been established if dynamic and/or static complexity are indeed relevant to engagement or learning, nor has it been established what would be the best conception or measure of these kinds of complexity if they are relevant. Therefore, this experiment examined dynamic and static complexity separately as an initial step to investigate which conceptions and measures of complexity (if any) were relevant to outcomes of engagement and learning. Combining static and dynamic as orthogonal complexity dimensions is beyond the scope of this study.
Because this experiment would measure learning, a between-subjects design was used (rather than repeated measures on the same participants) to avoid learning carrying over from the previous game to the next game to affect learning outcomes. To explain the manipulation of the level and type of complexity, first requires an explanation of the game created for this study.

### 6.3.1 Game Design

A similar approach was taken to study one above (see chapter, *Study 1: Evaluating Play Appeal Factors & Engagement*) at the broadest level, in that this study also employed an abstract game with adjustable parameters that could be objectively measured to establish a relationship between a game’s measurable properties and outcomes of engagement. But apart from this high-level similarity, many of the details were different, and so these differences will be explained in this section.

This second study required developing a completely new abstract game. The player would once again control an arrow on a two-dimensional world that wraps around, but almost everything else about the game was different. The game was built to have adjustable complexity in terms of both the sensitivity to initial conditions (to test the edge of chaos hypothesis), and in terms of the causal structure (to test the scale-free hypothesis). This meant splitting the game into two modes (one to test different levels of dynamic complexity, and one for testing levels of static complexity), which the game would automatically select on opening to try to balance the number of participants for the dynamic complexity mode, and the number for the static complexity mode, or otherwise choose one of them at random if they were already balanced in number.

The game mode for testing dynamic complexity (the edge of chaos hypothesis) consisted of a flocking algorithm (figure 6.2), chosen for its frequent mention as an example of emergence (e.g. Adams and Dormans, 2012; Berrondo and Sandoval, 2015; De Wolf and Holvoet, 2005; Silverberg et al., 2013; Sweetser, 2008; Szabo, Teo, and Chengleput, 2014). It consists of only three simple rules for each boid in the
flock, each rule forming a “pull” on the boid to help determine its current velocity (for details, see above citations):

1. Cohesion: Move towards the average position of all boids (long range attraction).

2. Separation: Move away from neighbours who are too close (short range repulsion).

3. Alignment: Move in the general direction that most of the flock is moving (matching average velocity).

The player’s avatar appeared and functioned as a larger version of the normal boids in the flock. The player controlled the avatar directly with the mouse (it did not respond to the other boids), but the boids responded to the avatar as a particularly powerful boid, to allow the player a means to intervention to learn about the flocking system through play. Dynamic complexity was varied by adjusting the sensitivity of the flocking algorithm (the strength of the three rules in determining current boid velocity) and consequently sensitivity to initial conditions, confirmed by measuring the Lyapunov exponent (see Calculating Lyapunov Exponent section below).
The game mode for testing static complexity (the scale-free hypothesis) consisted of a flat two-dimensional world containing five different kinds of gem, each occurring in two separate instances, to make for a total of 10 individual gems to be found in the game world. The player avatar was an arrow controlled by the mouse in the same manner as the dynamic complexity game mode, but with the additional ability to press the space bar or click the mouse to “grab” a nearby gem within a certain radius, then press it again to drop the gem, to allow the player to arrange the gems as desired (figure 6.3).
Figure 6.3: Screenshot of the game mode for testing static complexity levels using gems that activate each other (glow) when in proximity. The player-controlled arrow can pick up and move the gems to see which ones activate which others.

The gems and the player were connected in a causal network where each node was a kind of gem (or the player) and each link was the capacity for the source gem to activate the recipient gem (figure 6.4). Activation was shown visually as a gem lighting up and glowing. A gem can only activate another gem if it is currently active, with the exception of the player avatar, which is permanently active and thus can serve as the original cause of activation cascades. Links can form loops on the network, or a node can be connected to itself, meaning that activations of two or more of this kind of gem next to each other can be self-sustaining once the player starts it off.

Figure 6.4: Example of a causal network in the static complexity game mode. Arrows indicate which gems will activate which other gems when close-by. The player’s avatar (the white arrow on the right-hand side) can also activate gems.

The low complexity condition had a star topology (a low number of links in the network), the medium (emergent) complexity condition had a scale-free topology (a medium number of links), and the high complexity condition had a complete...
topology (the maximum number of possible links). Differing levels of static complexity were confirmed by measuring total walk count.

6.3.2 Checking Levels of Static & Dynamic Complexity

6.3.2.1 Dynamic Complexity Game

The inertia of the boids was the variable manipulated to create the difference between the conditions. The specific values used for each condition were tuned based on both informal pilot tests and Lyapunov calculations.

- In the **low complexity** (order) condition, the boids had very little inertia and responded almost instantly to flocking influences (figure 6.5). Thus, the overall state of the system would not be heavily dependent on initial conditions (but would inevitably be slightly dependent). This should create an orderly system that is simple to understand and control.

- In the **mid complexity** (emergence) condition, the boids had a moderate amount of inertia to give them a moderate but lagging response to flocking influences (figure 6.6). This would create a moderate degree of sensitivity to initial conditions owing to inertial differences. A medium degree of sensitivity to initial conditions should result in an emergent system according to the edge of chaos hypothesis, meaning complex behaviour should arise from the simple rules.

- In the **high complexity** (chaos) condition, the boids had a very large amount of inertia to give them a very delayed response to flocking influences (figure 6.7). This would create a very high degree of sensitivity to initial conditions. This should result in a chaotic system, with behaviour that seems close to or indistinguishable from random noise.
Figure 6.5: Screenshot of a typical flock formation produced from the low complexity condition. Boids would rapidly organise into a roughly pentagonal formation and keep their distances very consistently.

Figure 6.6: Screenshot of a typical flock formation produced from the medium complexity condition. Boids would gradually coalesce into a loose, flowing cluster, often passing between or around each other.
The dynamic complexity manipulations were confirmed by measuring the Lyapunov exponent. This involved simulating the system and calculating how much similar initial conditions either converge or diverge over time. The resulting distance calculations are not yet a Lyapunov exponent, only a Lyapunov number. Taking the natural log of a Lyapunov number will produce the Lyapunov exponent. If the exponent is negative, the states are converging, which is a sign of order and stability. If it is exactly zero, then they are neither converging nor diverging, but remaining the same distance apart, which is itself also a sign of stability. If it is greater than zero, then they are diverging, which is a sign of sensitivity to initial conditions and therefore a sign of chaos.
6.3.2.2 Calculating Lyapunov Exponent

The Lyapunov exponent is an established mathematical concept, describing the sensitivity to initial conditions as the exponential rate of divergence of selected nearby conditions in a system (see above literature review, page 66. But a full understanding is not necessary to understand this research).

It requires choosing initial conditions, then simulating the system until it seems to have settled into a relatively typical behaviour, for example, settling into an orbit on a strange attractor (this is to prevent accidentally choosing initial conditions that are actually very unlikely, anomalous, or unstable). Now that the system is in its normal rhythm, one records the current system state in an n-dimensional space that fully describes the system, and selects a different, arbitrarily close point on that space at a known distance (a very similar initial condition). Both initial points are then simulated forward in time, and then the distance between them in that n-dimensional space is re-calculated. If the distance between them has shrunk over time, then they are converging to a similar system state, which is a sign of stability and order. If the distance between them has increased, then they are diverging and becoming less and less similar over time, which is a sign of chaos. Ideally, this is done multiple times to get an average of how much that distance changes depending on where the system state is in the n-dimensional system space.

This process entails multiple judgment calls: Deciding when the system has settled into typical behaviour; the distance by which to separate the initial similar conditions; how far forward in time to simulate before checking if the similar points have become closer or further apart; how many times to go through this cycle before choosing a new starting point; how many times to repeat this entire process again by choosing new starting points.

To confirm the amount of chaos in each condition, I ran the game through 300 automated simulations of the system (100 for each complexity level) while recording data about the state of the system, getting the software to automatically simulate 24 variations on initial conditions (one variation for every dimension required to fully describe the system state: six boids with 4 dimensions each. Each boid has an
x and y position, and also an x and y velocity) every 40 timesteps, simulating ahead 100 timesteps before checking how much a variation deviated from the initial conditions. The recordings were started after several seconds when the system appeared to have settled into normal behaviour (see figures 6.5, 6.6, and 6.7). The resulting data allowed calculation of the Lyapunov exponent for each of the conditions – low, medium, and high degrees of chaos. The Lyapunov exponent is a measure of sensitivity to initial conditions, which is a common component of the definition of chaos (Cross, 2000; Habib and Ryne, 1995; Rosenstein, Collins, and De Luca, 1993).

Using the system’s state in its full n-dimensional space (24 dimensions, in the case of this system) works well for simple chaotic systems in the mathematics and system dynamics literature. However, it may not be appropriate for a flocking algorithm. Depending on the number of boids in the flock, the n-dimensional space describing the system in its totality could be extremely hyper-dimensional. This almost guarantees divergence, where system states are very far apart in the n-dimensional space, and yet are actually very similar.

For example, a flocking algorithm could be built to yield very stable, rigid behaviour where the boids lock into an equidistant grid formation every single time, rather than flowing around and past each other. Keep in mind that all the boids are identical in nature and behaviour, such that if one were to take boid number 15, and swap its position and velocity with boid number 6, there would be no distinguishable effect on the system - it would appear and behave exactly as though the two boids had not been switched. However, in the n-dimensional space of all possible system configurations, that switch would create a huge leap from one possible state, to another possible state, because the positions and velocities of each boid are independent dimensions. Therefore, this naive method of representing a flocking system state with separate dimensions for each boid is all but guaranteed to find divergence in the flocking system, even if they tend to form grid formation that is extremely ordered and stable. Similarly, the entire flock could be moved a large distance to the left and that would not significantly change its behaviour (merely displace it), and yet it would once again cause a significant leap in the n-dimensional possibility space.
Therefore, when it came to measure the amount of divergence in the system state, I used a different representation of the system state that was designed to minimise the above problems. It contained two dimensions per boid: One for the boid’s distance from the average centre of the entire flock, and one for the the distance between the boid’s velocity and the average velocity of the entire flock. This was a representation of flock formation that would not be affected by a mere displacement of the whole flock to another location in space. But it also needed to account for the fact that boids are interchangeable and that a switching of boid A with boid B is not a significant state change. Therefore, these two lists (boid average position distance, and boid average velocity distance) were sorted into ascending order. This ensured that if, for example, boids A and B switched locations, that change would not affect this representation of the system state (more technical details can be found in appendix H).

Other mathematical representations of a flocking system’s state may be developed in future that better capture how humans recognise and process flocking patterns and therefore what it is exactly that humans find so interesting about flocking behaviour, but such an investigation is outside the scope of this research.

Note that a full understanding of how to calculate the Lyapunov exponent is not necessary to understand the results of this study. For further explanation of the Lyapunov exponent, see literature review above, page 66. For more technical details of specifically how it was calculated with the flocking system of this study, see appendix H.

Now one must consider what Lyapunov exponent would be at the edge of chaos. Intuitively, one might assume it is zero, but such a system could actually be quite stable. Lacking any good reason to hypothesise a specific value for the edge of chaos, I decided to instead tune the game system to seem subjectively interesting, then measure that value, and compare it to values for the system when it was tuned to seem very ordered and when tuned to seem very chaotic. These are the values reported in this study. What matters is whether the orderly condition is the low-
est value, the chaotic condition the highest value, and the edge of chaos condition distinctly between the two extremes. This was accomplished, as shown in table 6.2.

<table>
<thead>
<tr>
<th>Lyapunov Exponent (n = 100)</th>
<th>Condition (Dynamic Complexity)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Lyapunov number</td>
<td>Mean 1.56</td>
</tr>
<tr>
<td></td>
<td>SD 0.76</td>
</tr>
<tr>
<td>Lyapunov exponent</td>
<td>Mean 0.36</td>
</tr>
<tr>
<td></td>
<td>SD 0.37</td>
</tr>
</tbody>
</table>

Table 6.2: The Lyapunov exponent measures for each of the three conditions in this study, used to verify level of dynamic complexity. The Lyapunov exponent is the natural log of the Lyapunov number. The larger the number, the greater the system’s sensitivity to initial conditions - a defining characteristic of chaotic systems.

6.3.2.3 Static Complexity Game

Different levels of static complexity were achieved by changing the number of causal interactions between game objects (which gems activated which other gems), thus altering the causal network of the game system. The number of gem types, differentiated by both shape and colour, was always five across all conditions. The player’s avatar also had a place in the causal network as it was able to activate gems by proximity, making for a total of six nodes in the network (five gems plus one player). The number of links varied by complexity level out of necessity.

To generate a network with similar properties as those discussed in the literature review as scale-free, the game followed a similar general procedure described in that literature for building the network in a rich-get-richer manner, whereby nodes with links are more likely to get yet more links (technical details can be found in appendix G). First, a random starting node was selected, then a random desti-

---

2An orderly system should have a Lyapunov exponent equal to or lesser than zero. The Lyapunov exponent of the low dynamic complexity condition was a positive number. Technically, this is a sign of chaos, meaning that the low complexity condition was not "orderly", at least with respect to this particular measure of sensitivity to initial conditions. This could be due to the method used to represent the system state described earlier: It attempted to minimise its sensitivity to trivial differences (e.g. shifting the entire flock to the left, or switching birds A and B), but it may not have completely succeeded (i.e., there may be better methods to represent flocking system states for measuring their Lyapunov exponent for studying human cognition and engagement). Alternatively, flocking systems themselves might be structurally predisposed to sensitivity to initial conditions, making it difficult to create a truly ordered system from a flocking algorithm. In any case, this will not hinder this study in investigating if there is a relationship between dynamic complexity and outcomes of engagement or learning.
nation node (with the player node excluded as a destination because it is already permanently active). A directed link was added to connect the starting node to the destination node.

From this point on, whenever a node is randomly selected it is (unless otherwise specified) selected indirectly via the list of links: i.e., instead of directly randomly selecting a node, it randomly selects a link from the list of links, then randomly selects one of the nodes to which that link is connected. This is to create the rich-get-richer effect of nodes with links being likely to accumulate yet more links. However, after the first link was added, only two nodes are connected by a link and therefore this selection process would only be able to select those two nodes that are already connected. Therefore, the selection of a link from the list was tweaked to instead randomly select from the size of the list plus one, where the selection of the extra item beyond the bounds of the list of links would instead trigger a direct selection of a random node from the list of nodes. This allowed the system to select nodes which were not yet connected by links, and the probability of doing so would shrink as the number of links grew: When there is only one link in the network, the chance of directly selecting a random node is 1 in 2; When there are two links, the chance is 1 in 3; When there are three links, 1 in 4; etc.

To ensure that there were no orphan nodes, and therefore that all gems could be activated to function usefully in the game, it then stepped through the list of nodes twice. The first time, it added a link leading from the node in question (the link going to another node, selected by the described link-based random selection process), and the second time around, it added a link leading to the node in question (the link leading to another node, selected by the link-based random selection process). This ensured all gem types would interact with at least one other gem type in the game.

Note that whenever a link was to be added, it would first check if there wasn’t already a link connecting the same two nodes in the same direction (i.e. A to B, or B to A). If the nodes were already connected in the same direction, then it would do nothing. This resulted in variability in the total number of links in the network.
once it was completed (but only for the scale-free condition. The star topology and completely connected conditions had no variability in their networks).

To increase the overall number of links in the network (to further differentiate this topology from the star and complete topologies), it would then attempt to add another eight links, for which the starting node and the destination node were both selected by the link-based random selection process. Once again, it would do nothing if a link already existed in the place selected.

Finally, it added a link starting from the player node and going to another node (selected by the link-based random selection process) in order to ensure that the player could directly interact with the network to start off the chain of activations.

- In the low complexity (low connectivity) condition, the causal network was a star topology, where a random gem type (not the player) was selected as the centre of the star, to which all other gem types were connected with bidirectional links. This should form a simple causal system in terms of both the number of linkages being low, and the overall structure being simple. The total number of links in the star was 11 (a link to the player was not necessary as it is always active), and total walk count (TWC) for the star topology was 41.

- In the mid complexity (emergent) condition, the causal network was built using a rich-get-richer algorithm, as is often used to produce scale-free networks. This should result in an emergent system, with a medium number of linkages in the system, and a complex but not completely random causal structure. Because this network was generated with an element of randomness, it had variation where the other two conditions did not. Therefore, I report the average and standard deviation of the TWC. The average number of links was 14.32 with a standard deviation of 1.12, and the TWC for the scale-free networks was an average of 97.54 with a standard deviation of 39.28 with a sample size of 1000 networks generated with the scale-free algorithm described above.

- In the high complexity (high connectivity) condition, the causal network was completely connected, such that every possible linkage was present. This should form a system that is complex in terms of the sheer number of discrete linkages
present, but is ultimately simple due to its maximally predictable and orderly structure. Total number of links was 30, and TWC for the maximum connectivity network was 970.

6.3.3 Measures

The game was entirely self-contained, including the introduction to the experiment to obtain informed consent, initial demographic questions, all play sections (initial free play, timed free play, and goal-oriented performance), the PANAS (see appendix D) and PENS questions (see appendix E), the explicit knowledge test, and the debriefing and option to withdraw from the study.

6.3.3.1 Dynamic Complexity Game

Both implicit and explicit knowledge were normalised as per the process defined in MMT’s formula 3.10 ($\rho_g = \frac{\hat{\rho}_g}{\rho_0}$): By taking the raw performance score and dividing it by the optimal performance to give a measure of overall mastery.

Explicit knowledge was tested by a screen asking “what are the rules determining how the creatures move?” Beneath were three text boxes where players could type answers. These text responses needed to be converted into numerical values for statistical analysis (i.e. "coding"). Two independent coders who were unaware of the hypotheses rated all participants’ answers for their accuracy in describing the actual rules of the flocking algorithm (the instructions given to the coders can be found in appendix F), with an inter-rater agreement rate of 84.96%. Disagreements were resolved by discussion. Scores were normalised by dividing by the maximum possible score (two points per rule, and three rules made for a maximum possible score of six).

The implicit knowledge test gave the player the goal of aligning the overall direction of the flock with a slowly, randomly rotating arrow in the centre of the screen, which would change from red to green depending on how well aligned the flock was to the direction of the arrow. There was also a score counter that counted
up in integers, the speed with which it ticked over to the next integer being deter-
mined by how well the player was doing at aligning the flock – when they were
pointed in the opposite direction, it would not count up at all; but when they were
all perfectly aligned with the arrow it would count up very rapidly. The text asked
the player to click to start the test. At which point, the score counter would activate
and a bar on the screen would start to drain down, indicating how much time was
left for this section. The implicit knowledge scores were normalised using the high-
est score achieved by an AI playing the game optimally, over 462 simulated runs
total (154 per complexity level).

6.3.3.2 Static Complexity Game

Implicit and explicit knowledge of the static complexity game were also normalised
as per the process defined in MMT’s formula 3.10 ($\rho_g = \frac{\rho_g}{\bar{o}_g}$). The implicit knowledge
test randomly redistributed the gems around the game world then presented partic-
ipants with the goal of activating a specific gem type, selected at random. Once they
had done so, a score counter ticked up another integer, and “Well done” was flashed
onscreen. Then the gems were randomly redistributed again and another gem was
selected for them to activate. A player would perform better at this task if they took a
direct route through the causal network and did not pick up any unnecessary gems.
This would allow them to complete each activation quicker and build up a higher
score. There was also a bar draining down to indicate how much time was left for
this section. As discussed in the section defining formula 3.10, a simple artificial in-
telligence (AI) was created to play the game perfectly (always going directly to the
gem/s that needed to be used to reach the goal), and its average score across $n=1000$
runs was used as $o_g$ in formula 3.10 to normalise the score of each participant.

The explicit knowledge test was a structured mental mapping task, where
participants were shown all the gem types laid out in a circle with the player avatar
as well, and asked to draw connections to indicate which gems activated which. This
mental model was compared to the game’s actual causal network (thus, the game’s
actual causal network - an explicit knowledge score of 100% - was used as \( o_g \) to give a normalised explicit knowledge score.

6.3.3.3 Engagement

Engagement was measured by triangulating three measures: The PENS questionnaire measuring game motivation (see appendix E), the PANAS tool to measure mood (see appendix D), and a free play period that measured how long participants chose to voluntarily interact with the game when there was no need to because it claimed to be “loading the next section”. Therefore, a total of two explicit measures and one implicit measure of engagement were given equal weight in calculating an overall engagement score for each participant.

The game was entirely self-contained, including the introduction to the experiment to obtain informed consent, initial demographic questions, all play sections (initial free play, timed free play, and goal-oriented performance), the PANAS and PENS questions, the explicit knowledge test, and the debriefing and option to withdraw from the study.
### Variable Definitions & Measures

<table>
<thead>
<tr>
<th>Variable</th>
<th>Source</th>
<th>Definition</th>
<th>Measure</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement</td>
<td>Triangulated from PANAS, PENS, and Free-Time Engaged</td>
<td>Participant’s level of engagement with a version of the game they played</td>
<td>Average of PANAS, PENS, and Free Time Engaged scores, normalised</td>
<td>0-1</td>
</tr>
<tr>
<td>PANAS</td>
<td>Watson and Vaidya, 2003</td>
<td>Participant’s self-rated mood after playing a version of the game briefly</td>
<td>Twenty Likert statement items, normalised</td>
<td>0-1</td>
</tr>
<tr>
<td>PENS</td>
<td>Rigby and Ryan, 2007; Przybylski, Rigby, and Ryan, 2010</td>
<td>Participant’s rating of game’s quality after playing a version of the game briefly</td>
<td>Nine Likert statement items, normalised</td>
<td>0-1</td>
</tr>
<tr>
<td>Free Time Engaged</td>
<td>Cameron, Banko, and Pierce, 2001; Lieberoth, 2015</td>
<td>Amount of time participant voluntarily spent playing a version of the game</td>
<td>Proportion of free-time period spent interacting with the game</td>
<td>0-1</td>
</tr>
<tr>
<td>Implicit Knowledge</td>
<td>Greiff et al., 2015</td>
<td>Ability to control and manipulate game system via direct, real-time interaction</td>
<td>Participant’s game score when assigned a goal to achieve, out of maximum possible score</td>
<td>0-1</td>
</tr>
<tr>
<td>Explicit Knowledge</td>
<td>Greiff and Fischer, 2013; Greiff et al., 2015</td>
<td>Accuracy of participant’s conscious understanding of game system</td>
<td>Dynamic Complexity</td>
<td>0-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity Static Complexity</td>
<td>Accuracy of participant’s description of game rules, as rated by blind coders</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Correct number of causal connections identified by participant, minus connections</td>
<td>-1-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>they failed to identify, or incorrectly identified, out of total number of actual connections</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3: Definitions and measures for dependent variables in experiment 2.1.


6.3.4 Procedure

Participants were gathered by sharing on social media and a university email newsletter a link to a webpage containing and browser-based version of the game (n = 222), and by approaching students in person at the university to play an offline version on a laptop (n = 44), with a total of 224 participants after removing outliers and blank responses (Ethics information in appendix C).

PANAS samples were discarded if they took less than the lower 5th percentile (13 seconds), or they did not select any responses. This resulted in 6 deletions (2 based on time, 4 based on inaction) for the dynamic complexity game, and 5 deletions (2 based on time, 3 based on inaction) for the static complexity game.

PENS samples were discarded if they took less than the lower 5th percentile (4 seconds. excluding one outlier who took 12 minutes), or they did not select any responses. This resulted in 5 deletions (1 based on time, 4 based on inaction) for the dynamic complexity game, and 6 deletions (3 based on time, 3 based on inaction) for the static complexity game.

Explicit knowledge samples in the dynamic complexity game were discarded if they took less than the lower 5th percentile (7 seconds). This resulted in 1 deletion. Samples that were left blank were kept because discarding these would discard participants who legitimately did not know what the correct answer was, but answers that were random keystrokes or entirely unrelated comments were discarded. This resulted in 2 deletions.

Explicit knowledge samples in the static complexity game were discarded if they took less than the lower 5th percentile (14 seconds). This resulted in 1 deletion. (Samples that did not do anything were kept because discarding these would discard participants who legitimately did not know what the correct answer was). One high outlier was also removed from medium complexity condition of the static complexity game.

Implicit knowledge samples in the static complexity game contained one very low scoring outlier in the low complexity condition.
One high-scoring outlier was also removed from the free-time samples in the medium complexity condition of the static complexity game.

For those with demographic data, there were 108 female, 107 male, and 9 NA, with a mean age of 26.86, and median age of 22 (table 6.4).

<table>
<thead>
<tr>
<th>Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity Type</td>
</tr>
<tr>
<td>Dynamic</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Static</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Total (Both Types)</td>
</tr>
</tbody>
</table>

Table 6.4: Demographics for experiment 2.1.

Offline participants were asked if they were interested in playing a short game on a laptop for approximately 10 minutes for a psychology study. Participants were left to play the game undisturbed as the researcher stood or sat off to one side and kept busy on their own device until the participant announced they had finished, at which point the participant was thanked and asked if they had any questions before they left (the offline version included the same written debriefing text as the online version, and consequently very few had any questions).

The researchers posted an open invitation to participate in a short 10-minute experiment on social media, and the invitation was distributed via the university’s research volunteering newsletter.

The game opened with introduction text to ensure informed consent, explaining that participation was voluntary and anonymous. Once participants had consented, they were asked to enter their age and gender if they wanted.

The next section was an initial free play period of 30 seconds where participants could interact freely with the game to pursue their curiosity. Text at the top of the screen explained that the mouse moved the player avatar around the game
world. The scale-free static complexity mode game had additional text to explain that some gems could activate other gems, and that click or pressing space bar will pick up a gem, and pressing it again will drop it. This first play section also would start with two gems right next to the player that should instantly light up, in order to make this causal possibility as immediately clear as the movement of the boids in the chaos dynamic complexity game mode.

Then participants were presented with PANAS and PENS in counterbalanced order. Based on pilot tests, the intent was to present these engagement scales when curiosity and engagement should be escalating or hitting its peak, rather than on its decline. This was followed by the final engagement measure: A free time section of 150 seconds where the game resumed, but with the addition of text reading, “The game is loading the next section of the experiment. Feel free to do what you like until it is ready,” along with a progress bar to show how much time was left. There was actually no loading going on, it was just a deception to ensure players did not feel pressured to interact with the game during that time. This deception was explained in the debriefing text at the end of the experiment. It was possible to exceed 150 seconds in this free play section because it would only continue when the player pressed the continue button, therefore some engagement times are over 150 seconds.

Based on pilot tests, this free time should be long enough for participants to satisfy their curiosity. Therefore, they will have learned all that they would if this were an educational sandbox game and they were not being forced to play it. The next sections were the implicit and explicit knowledge tests, in counterbalanced order. The explicit knowledge test’s duration was entirely dependent on how long the participant chose to take on that section, but the implicit knowledge test lasted 30 seconds for the edge of chaos game mode and 60 seconds for the scale-free game mode, based on pilot tests.

Finally, participants were presented with a debriefing screen where they were thanked, the deception of the “free time” section explained, and given the option to withdraw from the study.
6.4 Results: Experiment 2.1

6.4.1 Dynamic Complexity Game

A two-way ANOVA found no significant difference between the groups in terms of engagement, $F(2, 109) = 2.38$, ($p = .097$) or explicit learning, $F(2, 86) = 1.779$, ($p = .175$), but there was a significant difference for implicit learning, $F(2, 93) = 9.811$, ($p < .001$). Post-hoc tests with Bonferroni correction found no support for the hypothesis that implicit learning would be higher in the medium complexity than the high complexity condition, $t(93) = .878$, ($p = 1$), but found that implicit learning in the low complexity condition was significantly higher than both the medium, $t(93) = 4.203$, ($p < .001$) and high complexity conditions, $t(93) = 3.471$ ($p = .002$).

**Pairwise Comparisons: Dynamic Complexity**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comparison</th>
<th>Mean Difference (I-J)</th>
<th>St. Error</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement</td>
<td>Low (I) – Medium (J)</td>
<td>0.01</td>
<td>0.03</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Low (I) – High (J)</td>
<td>0.06</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Medium (I) – High (J)</td>
<td>0.06</td>
<td>0.03</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Low (I) – Medium (J)</td>
<td>0.24</td>
<td>0.06</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Implicit Knowledge</td>
<td>Low (I) – High (J)</td>
<td>0.20</td>
<td>0.06</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>Medium (I) – High (J)</td>
<td>-0.05</td>
<td>0.05</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Low (I) – Medium (J)</td>
<td>0.03</td>
<td>0.05</td>
<td>1.00</td>
</tr>
<tr>
<td>Explicit Knowledge</td>
<td>Low (I) – High (J)</td>
<td>0.09</td>
<td>0.05</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Medium (I) – High (J)</td>
<td>0.06</td>
<td>0.05</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 6.5: Pairwise comparisons for the dynamic complexity game in experiment 2.1.

Figure 6.8: Mean engagement, implicit and explicit knowledge for low, medium, and high conditions of dynamic complexity, shown with 95% confidence intervals.
6.4.2 Static Complexity Game

A two-way ANOVA found no significant difference between the groups in terms of engagement, $F(2, 107) = 1.347$, ($p = .264$) or explicit learning, $F(2, 96) = 1.089$ ($p = .341$), but there was a significant difference for implicit learning, $F(2, 74) = 14.858$ ($p < .001$). Post-hoc tests with Bonferroni correction found that implicit learning was significantly lower in the medium complexity condition in comparison to both the low complexity condition, $t(74) = -2.598$, ($p = .034$), and the high complexity condition, $t(74) = -5.449$, ($p < .001$).

Descriptive statistics for experiment 2.1 are shown in table 6.7.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comparison</th>
<th>Mean Difference (I-J)</th>
<th>St. Error</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engagement</td>
<td>Low (I) – Medium (J)</td>
<td>0.03</td>
<td>0.03</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Low (I) – High (J)</td>
<td>0.05</td>
<td>0.03</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Medium (I) – High (J)</td>
<td>0.02</td>
<td>0.03</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Low (I) – Medium (J)</td>
<td>0.34</td>
<td>0.13</td>
<td>0.03</td>
</tr>
<tr>
<td>Implicit Knowledge</td>
<td>Low (I) – High (J)</td>
<td>-0.25</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Medium (I) – High (J)</td>
<td>-0.56</td>
<td>0.11</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Low (I) – Medium (J)</td>
<td>0.05</td>
<td>0.05</td>
<td>0.83</td>
</tr>
<tr>
<td>Explicit Knowledge</td>
<td>Low (I) – High (J)</td>
<td>0.07</td>
<td>0.05</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>Medium (I) – High (J)</td>
<td>0.01</td>
<td>0.04</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 6.6: Pairwise comparisons for the static complexity game in experiment 2.1.

Figure 6.9: Mean engagement, implicit and explicit knowledge for low, medium, and high conditions of static complexity, shown with 95% confidence intervals.
<table>
<thead>
<tr>
<th>Complexity Type</th>
<th>Level</th>
<th>Dependent Variable</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic</td>
<td>Low</td>
<td>Engagement</td>
<td>35</td>
<td>0.61</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Explicit Knowledge</td>
<td>28</td>
<td>0.24</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Implicit Knowledge</td>
<td>27</td>
<td>0.27</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>Engagement</td>
<td>37</td>
<td>0.60</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Explicit Knowledge</td>
<td>30</td>
<td>0.24</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Implicit Knowledge</td>
<td>30</td>
<td>0.24</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Engagement</td>
<td>37</td>
<td>0.60</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Explicit Knowledge</td>
<td>36</td>
<td>0.04</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Implicit Knowledge</td>
<td>32</td>
<td>0.18</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Engagement</td>
<td>31</td>
<td>0.64</td>
<td>0.14</td>
</tr>
<tr>
<td>Static</td>
<td>Medium</td>
<td>Explicit Knowledge</td>
<td>35</td>
<td>0.01</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Implicit Knowledge</td>
<td>24</td>
<td>0.30</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Engagement</td>
<td>39</td>
<td>0.58</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Explicit Knowledge</td>
<td>36</td>
<td>0.00</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Implicit Knowledge</td>
<td>36</td>
<td>0.89</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Table 6.7: Descriptive statistics for experiment 2.1.

6.5 Discussion of Experiment 2.1

These results support H3 that game complexity influences learning of that game system, although differences were only significant for implicit learning, not explicit. The hypothesis that game complexity influences engagement (H1) was not supported in this experiment. A summary of which hypotheses were supported or not are shown in table 6.8.
Table 6.8: Hypotheses and findings for experiment 2.1.

Apart from H1 and H3, there were several specific predictions that could be extracted from the theory in the literature on game design, motivational psychology, cognitive science, and complexity theory. However, these predictions were not supported. The Inherently Learnable Systems hypothesis postulated that increased player engagement would result in improved learning. The Emergent Gameplay hypothesis extracted from the literature postulated that an intermediate degree of complexity would result in the most interesting game and would thus result in the most engagement. This prediction was also concordant with the theories of flow, and curiosity as incongruity or optimal arousal. These theories were not supported in this experiment.
STUDY 2: GAME COMPLEXITY & CURIOSITY

The Epistemic Emergence hypothesis postulated that the simple rules giving rise to complex behaviour would make the underlying rules difficult to infer from the complex behaviour, and therefore mastering the game through hands-on play would not guarantee explicit understanding of those underlying rules. This should have resulted in high implicit learning but low explicit learning in the emergent (medium complexity) conditions. Such an effect was not observed in this experiment. Implicit learning scores for the low dynamic complexity condition were higher than the other two conditions. Implicit learning scores for medium static complexity condition were lower than the other two conditions. The different pattern of results between the static and dynamic forms of complexity support the idea that multiple kinds of complexity may be necessary to fully model and understand engagement and learning in complex games.

The Edge of Chaos hypothesis postulated that a medium level of dynamic complexity would result in emergence, and therefore (according to Emergent Gameplay and Epistemic Emergence) high engagement, high implicit learning, but low explicit learning. The Scale-Free Network hypothesis postulated that a medium level of static complexity (specifically, a scale-free structure to the causal network of the game system) would also result in emergence, and therefore also cause the results of the Emergent Gameplay and Epistemic Emergence Hypotheses. The results did not support these theories of what constitutes static or dynamic emergence, but the different implicit learning results for static vs. dynamic complexity suggests it is an important distinction to make in matters of cognition. A full discussion of these theoretical implications will follow after presenting the results of the second experiment.

It is interesting to note that the explicit learning results in the static complexity conditions show a pattern that is more consistent with a naive conception of static complexity (where complete connectivity is high complexity), whereas the implicit learning results in the static complexity conditions show a pattern that is more consistent with an information theory perspective of static complexity (where complete connectivity is low complexity). This could suggest that at an implicit level, participants treated the completely connected system in a simple and efficient man-
uder consistent with an information theory conception of static complexity (wherein a completely connected system can be easily represented with relatively little data), but at an explicit level, participants reverted to a naive conception of static complexity and the sheer number of connections was overwhelming. This could indicate that different measures of static complexity might be appropriate for studying implicit versus explicit learning processes. The different possible interpretations based on a naive or information theoretic conception of static complexity will prove even more relevant after the second experiment.

Possible effects of engagement in the data did not reach significance, and so the procedure and measures were redesigned for a second experiment to more deeply explore how complexity may affect engagement.

The experiment was redesigned for Experiment 2.2 to increase its power to detect differences in engagement, switching to a within-subjects design that could directly compare levels of relative preference for each individual, and including a measure of trait curiosity as a theoretically relevant predictor of interest in complexity. This experiment will address H1 (Game complexity influences engagement) and H2 (Trait curiosity influences engagement), as defined above.

6.6 Method: Experiment 2.2

A within-subjects design was used to increase the experiment’s ability to detect differences between conditions in comparison to the previous experiment. Because learning was not a dependent variable in this next experiment, there was less concern for carry-over of learning effects that could occur for repeated-measures designs. But to minimise its effect on other dependent variables, the ordering in which participants experienced different conditions was randomised. The resulting design used complexity level as a within-groups factor, and complexity type (dynamic or static) as a between-groups factor (although the groups were not compared to each other, so all statistical tests were repeated-measures tests). Within both groups, each
participant would experience all three levels of complexity, but in a different, ran-
domised order.

**Conditions**

<table>
<thead>
<tr>
<th>Group 1: Dynamic Complexity</th>
<th>Group 2: Static Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Complexity</td>
<td>Low Complexity</td>
</tr>
<tr>
<td>Medium Complexity</td>
<td>Medium Complexity</td>
</tr>
<tr>
<td>High Complexity</td>
<td>High Complexity</td>
</tr>
</tbody>
</table>

Table 6.9: Repeated-measures design of experiment 2.2. Each participant experienced all three levels of complexity, in randomised order, for one type of complexity (static or dynamic). The two groups were not compared statistically, only used for organisation.

**6.6.1 Game Design**

The same game was used as in Experiment 2.1 (see above section 6.3.1).

**6.6.2 Measures**

Engagement was measured with two Likert-type items asking participants to rate the extent to which they agreed (on a five-point scale) with the statements, “I preferred this version of the game,” and, “I found this version of the game most interesting,” respectively. The measure of preference was combined with the measure of interest to create a general measure of engagement that was an average of the two.

Trait curiosity was measured using the Curiosity and Exploration Inventory Mark II (CEI-II), consisting of 10 Likert-type questions. These items were weighted according to their loading on the construct (see below section, *Evaluating the Curiosity Measure CEI-II*) to give a single trait curiosity score for each participant.

The amount of free time spent interacting with each game was measured as a control variable. All three versions of the game (low, medium, and high complexity) were presented simultaneously during the free time period, giving participants the mutually exclusive choice of which level of complexity they most wanted to play with. To help distinguish them visually, each of the three versions had a different background tile pattern (square, hexagonal, or cross patterns) and colour, counter-
balanced randomly (These different background textures can be seen in figures 6.5, 6.6, and 6.7).

The free time measure was the total amount of time the participant spent interacting with one version of the game, divided by the total amount of time the participant spent interacting with any of the different versions of the game, thereby giving a proportion of the total amount of time dedicated to play. For example, if participant A spent a total of 100 seconds interacting with all the different versions of the game in the free time period, and 20 of those seconds were spent interacting with the low complexity version, then they would have time score of 20% for low complexity. If participant B spent a total of only 40 seconds interacting with all the different versions, and only 8 of those seconds interacting with the low complexity version, then they would also have time score of 20% for low complexity. This procedure accommodates for individual differences in overall levels of interest in the game in order to capture individual differences due to the different complexity levels.

**Variable Definitions & Measures**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Source</th>
<th>Definition</th>
<th>Measure</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preference</td>
<td>Self-developed</td>
<td>How much participant said they prefer each level of complexity</td>
<td>Likert Statement</td>
<td>0-4</td>
</tr>
<tr>
<td>Interest</td>
<td>Self-developed</td>
<td>How interesting participant rated each level of complexity</td>
<td>Likert Statement</td>
<td>0-4</td>
</tr>
<tr>
<td>Engagement</td>
<td>Self-developed</td>
<td>Extent to which each complexity level tended to engage participant</td>
<td>Average of Preference and Interest, normalised</td>
<td>0-1</td>
</tr>
<tr>
<td>Curiosity</td>
<td>Grossnickle, 2016</td>
<td>Level of curiosity as a character trait in participant</td>
<td>CEI-II</td>
<td>0-1</td>
</tr>
<tr>
<td>Time</td>
<td>Cameron, Banko, and Pierce, 2001; Lieberoth, 2015</td>
<td>Amount of time participant spent interacting with the game</td>
<td>Free time interacting with each level of complexity, presented simultaneously</td>
<td>0-1</td>
</tr>
</tbody>
</table>

Table 6.10: Definitions and measures for dependent variables in experiment 2.2.
6.6.3 Procedure

Participants were gathered in two rounds of recruitment using the Mechanical Turk online research recruitment service, reimbursing each participant with $1.50US (Ethics information in appendix C). Two hundred and one samples were collected in the first round and 104 collected in the second, for a total of 305 participants (after removing outliers and blank responses). For those with demographic data, there were 70 female, 155 male, and 18 NA, with a mean age of 33.39, and median age of 31.

One participant opted to withdraw. Responses were discarded if they took less time than the lower fifth percentile to answer (i.e. less than seven seconds for the Likert engagement questions, or ten seconds for the CEI-II questions). This resulted in two deletions of engagement responses and eight deletions of CEI-II responses for the dynamic complexity game, and nine deletions of engagement responses and seven deletions of CEI-II responses for the static complexity game.

<table>
<thead>
<tr>
<th>Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Table 6.11: Demographics for experiment 2.2.

Total average experiment duration was seven minutes 32 seconds (median six minutes 34 seconds) for the static complexity version of the game, and three minutes 56 seconds (median three minutes 37 seconds) for the dynamic complexity game. For both static and dynamic versions, the average was five minutes 23 seconds (median four minutes 19 seconds).

Where Experiment 2.1 was a between-subjects design (each participant experienced a different level of complexity), Experiment 2.2 was a within-subjects design (each participant experienced all three levels of complexity - low, medium, and high - but only one type of complexity - dynamic or static).
As in Experiment 2.1, the game opened with introduction text to ensure informed consent, followed by entering demographic information of age and sex. Then an initial free play period presented each version of the game (low, medium, and high complexity), one after the other (30 seconds each for the dynamic complexity game, and 80 seconds each for the static complexity game). Which version was played first, second, or third was counterbalanced randomly.

Then participants were presented with all three versions (low, medium, and high complexity) simultaneously, next to each other horizontally across the screen. Which version was in the left, middle, or right position was counterbalanced randomly. This was the free time period in which participants could interact with any of the three versions of the game they wanted while a fake progress bar claimed to show that the next part of the game was loading, and the text, “You can play whichever version of the game you prefer”.

After this, participants were asked to rate their explicit degree of preference and interest in the three versions of the game using the two Likert-type items described above (the corresponding Likert item was shown over each version of the game – left, middle, or right – putting a total of six Likert items on this screen).

Then participants were presented with the CEI-II tool to measure their trait curiosity. This was presented after the game play and engagement measures in order to prevent the CEI-II from priming participants and having their thoughts on trait curiosity affect their measures for engagement.

Finally, a debriefing screen explained the experiment, provided contact information, and an opportunity to withdraw their data from the study.

6.6.4 Evaluating the Curiosity Measure CEI-II

Cronbach’s alpha was .886. One item would increase Cronbach’s alpha to .889 if deleted, but all the rest would decrease it if deleted. To retain consistency with the tools use in other literature, all items were kept.
Factor analysis suggested (judging by eigenvalues greater than one) CEI’s questions might measure at most 2 dimensions of curiosity. However, all items except one factored into both dimensions.

This was not a study specifically to evaluate ways to improve or modify CEI-II, and therefore it was used as it was designed. Weightings of the items on the final trait curiosity score for each participant (from confirmatory factor analysis) are shown in table 6.12.

<table>
<thead>
<tr>
<th>Curiosity &amp; Exploration Inventory Mark 2 Item Weightings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
</tr>
<tr>
<td>Weight</td>
</tr>
</tbody>
</table>

Table 6.12: Weightings of the items in CEI-II as calculated from confirmatory factor analysis.

6.7 Results: Experiment 2.2

6.7.1 Dynamic Complexity Game

A MANCOVA found no significant effect for complexity, Wilk’s Lambda = .31, F(4, 113) = 1.21, p = .31, \( \eta^2 = .04 \), or for complexity and curiosity (as measured with CEI-II), Wilk’s Lambda = .98, F(4, 113) = .516, p = .72, \( \eta^2 = .02 \). We thus fail to reject the null hypothesis and indeed there were no significant pairwise comparisons. The control variable of free time spent interacting was found to have no correlation or interaction with other variables and was thus dropped from the model.

![Figure 6.10: Mean preference, interest, and engagement for low, medium, and high conditions of dynamic complexity, shown with 95% confidence intervals.](image)
It is interesting to note that the low complexity condition (which resulted in the most implicit learning in Experiment 2.1) was also rated as the most engaging, although this difference was not statistically significant (explicit interest in low complexity compared to medium complexity, $p = .445$, and compared to high complexity, $p = 1$).

The hypothesis that curiosity would correlate with engagement with more complexity was not supported. Instead, the most statistically significant finding was that trait curiosity as measured with CEI-II correlated significantly with engagement in the low complexity condition, $r(123) = .181$ ($p = .047$). Many other correlations with curiosity achieved similar levels of significance: Preference for low complexity, $r(123) = .158$ ($p = 0.08$), interest for low complexity, $r(123) = .17$ ($p = 0.061$), medium complexity, $r(123) = .175$ ($p = 0.053$), and high complexity, $r(123) = .168$ ($p = 0.063$), and Engagement for high complexity, $r(120) = .178$ ($p = 0.052$).

Keeping in mind the MANCOVA and pairwise comparisons were not significant, the curiosity correlations were, and there are certain trends worth noting in the graphs of trait curiosity plotted against complexity level: There seemed to be a general preference for lower complexity, and this is strongest for those with the lowest trait curiosity. Those with higher trait curiosity seem to be more engaged than those with lower trait curiosity, and also less affected by complexity level – they are more interested regardless of how complex the game is (figure 6.11).
Figure 6.11: Engagement plotted against dynamic complexity level, with different quartiles of CEI-II score shown as separate curves. Shown trends: Low curiosity has decreasing engagement with higher complexity, but high curiosity has higher engagement across all complexity levels.

6.7.2 Static Complexity Game

A MANCOVA found a significant effect for complexity, Wilk’s Lambda = .143, F(4, 116) = 4.84, p = .001, $\eta^2 = .14$, and for complexity and curiosity (as measured with CEI-II), Wilk’s Lambda = .89, F(4, 116) = 3.619, p = .008, $\eta^2 = .11$. The control variable of free time interacting was found to have no correlation or interaction with other variables and was thus dropped from the model.

Pairwise comparisons using Bonferonni correction found significant results: The high complexity condition was rated as more interesting than the low complexity condition (p = 0.043); and the engagement score was higher for the high complexity condition than the medium complexity condition (p = 0.09).
Pairwise Comparisons

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comparison</th>
<th>Mean Difference (I-J)</th>
<th>St. Error</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preference</td>
<td>Low (I) – Medium (J)</td>
<td>-0.19</td>
<td>0.17</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>Low (I) – High (J)</td>
<td>-0.22</td>
<td>0.18</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Medium (I) – High (J)</td>
<td>-0.39</td>
<td>0.21</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Low (I) – Medium (J)</td>
<td>-0.24</td>
<td>0.18</td>
<td>0.59</td>
</tr>
<tr>
<td>Interest</td>
<td>Low (I) – High (J)</td>
<td>-0.45</td>
<td>0.18</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>Medium (I) – High (J)</td>
<td>-0.21</td>
<td>0.20</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>Low (I) – Medium (J)</td>
<td>-0.05</td>
<td>0.04</td>
<td>0.63</td>
</tr>
<tr>
<td>Engagement</td>
<td>Low (I) – High (J)</td>
<td>-0.08</td>
<td>0.04</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Medium (I) – High (J)</td>
<td>-0.03</td>
<td>0.05</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 6.13: Pairwise comparisons for the static complexity game in experiment 2.2.

The hypothesis that curiosity would correlate with engagement in high complexity was not supported, but it did correlate with engagement in medium complexity. Trait curiosity as measured with CEI-II correlated significantly with explicit preference for low complexity, r(123) = .385 (p < 0.001), and medium complexity, r(123) = .187 (p = .038), with explicit interest in low complexity, r(123) = .399 (p < .001), and medium complexity, r(123) = .328 (p < .001), and explicit engagement with low complexity, r(123) = .417 (p < .001), and medium complexity, r(123) = .28 (p = .002).

Inspecting the graphs plotting trait curiosity against complexity level reveals some interesting patterns. The high complexity condition seemed to engage players of all curiosity levels roughly equally. Again, it appears the more curious individuals are overall more interested than less curious individuals, regardless of
how complex the game is. But they do seem to show a slight preference for lower static complexity (i.e. less connectivity in the system).

Descriptive statistics are shown in table 6.14.

![Figure 6.13: Engagement plotted against static complexity, with CEI-II quartiles shown as separate curves. Shown trends: Low curiosity has increasing engagement with complexity, but high curiosity has decreasing engagement with complexity.](image)

### Table 6.14: Descriptive Statistics for experiment 2.2.

<table>
<thead>
<tr>
<th>Complexity Type</th>
<th>Level</th>
<th>Dependent Variable</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic</td>
<td>Low</td>
<td>Preference</td>
<td>2.20</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interest</td>
<td>2.35</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Engagement</td>
<td>0.57</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>Preference</td>
<td>2.17</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interest</td>
<td>2.16</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Engagement</td>
<td>0.54</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Preference</td>
<td>2.10</td>
<td>1.07</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Interest</td>
<td>2.21</td>
<td>1.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Engagement</td>
<td>0.54</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Preference</td>
<td>2.11</td>
<td>1.05</td>
</tr>
<tr>
<td>Static</td>
<td>Medium</td>
<td>Interest</td>
<td>2.26</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Engagement</td>
<td>0.55</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Preference</td>
<td>2.20</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Interest</td>
<td>2.28</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Engagement</td>
<td>0.56</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Preference</td>
<td>2.17</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>Interest</td>
<td>2.36</td>
<td>1.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Engagement</td>
<td>0.57</td>
<td>0.28</td>
</tr>
</tbody>
</table>
6.8 Discussion of Experiment 2.2

The results of this follow-up experiment support H1 that game system complexity influences engagement, but only for static complexity, and H2 that trait curiosity influences engagement with complexity. A summary of which hypotheses were supported or not is shown in table 6.15.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Description</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>Game system complexity influences engagement</td>
<td>Supported</td>
</tr>
<tr>
<td>H2</td>
<td>Trait curiosity influences engagement with complexity</td>
<td>Supported</td>
</tr>
<tr>
<td>Emergent Gameplay</td>
<td>Emergent (medium complexity) conditions will have higher engagement than low and high complexity conditions</td>
<td>Not Supported</td>
</tr>
<tr>
<td>/ Flow / Curiosity as Incongruity or Optimal Arousal Edge of Chaos</td>
<td>Medium dynamic complexity (the edge of chaos) condition will have higher engagement and implicit learning (but not explicit) than low and high dynamic complexity conditions</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Scale-Free Network</td>
<td>Medium static complexity (scale-free network) condition will have higher engagement and implicit learning (but not explicit) than low and high static complexity conditions</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Challenging Complexity / Drive Theory of Curiosity</td>
<td>High complexity conditions will have higher engagement than low and medium complexity conditions</td>
<td>Not Supported (Depending on definition of static complexity)</td>
</tr>
<tr>
<td>Effectance</td>
<td>Low complexity conditions will have higher engagement than medium and high complexity conditions</td>
<td>Supported (Depending on definition of static complexity)</td>
</tr>
<tr>
<td>Curious About Complexity</td>
<td>Individuals with higher trait curiosity will be more interested in higher levels of complexity</td>
<td>Not Supported</td>
</tr>
</tbody>
</table>

Table 6.15: Summary of the hypotheses and findings of experiment 2.2.

These results do not support the Emergent Gameplay hypothesis. But in combination with Experiment 2.1, do show some tentative support for the Inherently Learnable Systems hypothesis – positing that learning and engagement go hand-in-hand in game play, and that therefore the systems which players are able to teach
themselves are more appealing to players, or the games that players find appealing they are able to teach themselves (the direction of causality needs further investigation). For both static and dynamic complexity, the levels of complexity that resulted in the most implicit learning were also those rated as most interesting by players (although for dynamic complexity, that difference was not statistically significant).

The hypothesis that more curious individuals would be more engaged with more complex games was not supported. On the contrary, the pattern in the data was that individuals with more trait curiosity were more engaged regardless of complexity level, having no clear preference for dynamic complexity levels (conceptualised as sensitivity to initial conditions), and showing an apparent preference for less static complexity (conceptualised as total walk count). This could be explained in terms of the challenge of discovering causal linkages: A system with fewer causal links presents a greater challenge in finding them, and therefore engages more curious players to explore every possibility to find all of them. In contrast, individuals with less trait curiosity were more engaged with more static complexity (more connectivity in the system). This could be explained in terms of effectance motivation (Klimmt, Hartmann, and Frey, 2007): A more densely connected system has a higher chance of displaying a reaction to a random action from the player, and therefore, less curious players may be more satisfied with the immediate gratification of seeing something happen from almost every action they take.

The trends in the data suggest that the less curious individuals may be more sensitive to differences in dynamic complexity (though not statistically significant). Overall, such results support a focus on catering for the needs of the less curious individuals in game design, as it seems the more curious will be more engaged across all complexity levels. They also overall suggest that less dynamic complexity (less sensitivity to initial conditions) is considered more interesting than higher dynamic complexity, and that more static complexity (i.e. a completely connected causal network) is considered more interesting than low static complexity (e.g. a star topology). The complexity levels rated as most interesting in Experiment 2.2 were also the levels of complexity that resulted in the most implicit mastery in Experiment 2.1,
showing some initial support for the Inherently Learnable Systems hypothesis, but this concept requires further investigation.

These results reveal some intriguing directions for future research, and have interesting implications for the design of educational games.
### Results

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Description</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>Game system complexity influences engagement</td>
<td>Supported</td>
</tr>
<tr>
<td>H2</td>
<td>Trait curiosity influences engagement with complexity</td>
<td>Supported</td>
</tr>
<tr>
<td>H3</td>
<td>Game system complexity influences learning of that game system</td>
<td>Supported</td>
</tr>
<tr>
<td>Inherently Learnable Systems</td>
<td>Conditions that are engaging will result in learning</td>
<td>Partially Supported: Only implicit learning, not explicit</td>
</tr>
<tr>
<td>Epistemic Emergence</td>
<td>Emergent (medium complexity) conditions will have high implicit learning but low explicit learning</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Emergent Gameplay / Flow / Curiosity as Incongruity or Optimal Arousal</td>
<td>Emergent (medium complexity) conditions will have higher engagement than low and high complexity conditions</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Edge of Chaos</td>
<td>Medium dynamic complexity (the edge of chaos) condition will have higher engagement and implicit learning (but not explicit) than low and high dynamic complexity conditions</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Scale-Free Network</td>
<td>Medium static complexity (scale-free network) condition will have higher engagement and implicit learning (but not explicit) than low and high static complexity conditions</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Challenging Complexity / Drive Theory of Curiosity</td>
<td>High complexity conditions will have higher engagement than low and medium complexity conditions</td>
<td>Not Supported (Depending on definition of static complexity)</td>
</tr>
<tr>
<td>Effectance</td>
<td>Low complexity conditions will have higher engagement than medium and high complexity conditions</td>
<td>Supported (Depending on definition of static complexity)</td>
</tr>
<tr>
<td>Curious About Complexity</td>
<td>Individuals with higher trait curiosity will be more interested in higher levels of complexity</td>
<td>Not Supported</td>
</tr>
</tbody>
</table>

Table 6.16: Summary of the hypotheses and findings of study two.
6.9 Discussion

6.9.1 Practical Applications in Educational Game Design

Objectively quantifiable properties of game systems could prove invaluable in predicting outcomes of engagement and learning, and therefore could be vital in the game design process, but much additional work is needed to establish which measures and theories of complexity are appropriate and effective for these purposes. In this study, the Lyapunov exponent measuring the sensitivity to initial conditions of a flocking game proved relevant to implicit learning, but not explicit learning or engagement. The topology of the causal network of the game (measured using total walk count) proved relevant to implicit learning and engagement, but not explicit learning. This measure of static complexity proved useful in testing the level of complexity participants rated as most interesting, but how exactly the degree of static complexity is conceptualised and measured will affect which network topology will be rated as more or less complex (as detailed in the discussion above). Total walk count may be suitable for testing the scale-free network hypothesis of emergent gameplay, but average shortest path provided a much more insightful explanation of the results in terms of challenge. A wealth of other conceptions of complexity can be found in the complexity theory literature that may prove to be of central importance to engagement and learning in games.

The two experiments in this study had several findings relevant to the design of educational games: Game systems with less sensitivity to initial conditions (i.e. more order and less chaos) may enable better implicit learning. Game systems with a scale-free causal structure may be result in less implicit learning than game systems with low connectivity (i.e. a star topology) or complete connectivity in their causal structure. Complete connectivity was also considered by players to be more interesting than star and scale-free topologies (Future studies will need to test other topologies). These findings show partial support for the speculations of some authors that an inherently fun system is an inherently learnable system (the systems that were rated the most engaging were also the systems that resulted in the most implicit learning), but contradicts the Emergent Gameplay hypothesis that medium
levels of complexity would be where to find that ideal degree of complexity that leads to learning and engagement. Other system structures and definitions of complexity should be explored to confirm this.

For educational game design, these findings would support trying to make the game as ordered and as fully connected as possible while still achieving the educational goal. There are various considerations and compromises in modelling a real-world phenomenon for educational simulation (Sterman, 2000), and facilitating play via engagement and implicit learning is another factor that will have to be weighed in those decisions.

Trait curiosity is a significant variable in player interest in game complexity. In general, the results suggest that those with higher trait curiosity are more engaged regardless of the game’s complexity level (not supporting the Curious About Complexity hypothesis that highly curious players would be bored by the simple conditions and more interested in the complex conditions). In contrast, those with lower trait curiosity may be more sensitive to differences in dynamic complexity. For dynamic complexity, the results suggest that less sensitivity to initial conditions was generally considered more interesting to less curious players (although this trend was not statistically significant). For static complexity, more connectivity was considered more interesting to less curious players, but less connectivity was considered more interesting to more curious players (not significant). Therefore, highly (completely) connected game systems may be equally interesting to players regardless of their individual trait curiosity, whereas low connectivity game systems (e.g. star topology) may be more interesting to more curious players, and less interesting to less curious players. This suggests that less curious players are less keen to find rare causal connections, and more appreciative of having more interactions immediately available, which is consistent with the play as effectance theory: If they can easily make something happen, less curious players find that kind of game more interesting, possibly because it is more immediately satisfying to be able to affect the game with any interaction at all. Overall, these results support focussing game design decisions on those with lower curiosity, as those with higher curiosity will tend to be more interested across all complexity levels.
6.9.2 Theoretical Implications

The opposing trends in the findings suggest a qualitative difference between static and dynamic forms of complexity. Therefore, multiple measures of complexity may be necessary to assess how different kinds of complexity can differentially affect learning and engagement.

There needs to be much additional work establishing standardised measures of complexity for research into game learning and engagement.

The findings partially supported the conjecture by some authors that an inherently fun system is an inherently learnable system. That point where learnability and engagement overlap was hypothesised to be at an emergent degree of complexity hovering between simplicity and complexity. But learning and interest actually co-occurred at the lowest level of dynamic complexity and the highest level of static complexity tested in this research. This conflicts with the emergence hypotheses, but also highlights a neglected area in much need of future research - what conceptions of complexity are appropriate and effective for predicting engagement and/or learning in games? There needs to be much additional work establishing standardised measures of complexity for research into game learning and engagement.

It is interesting that Klarkowski et al., 2015 also attempted to measure flow across levels of low, medium, and high challenge, with results that also failed to match the inverted U-shape predicted by the theory (with low and medium challenge resulting in very similarly high flow outcomes, but high challenge resulting in low flow).

There are several possible interpretations of our results regarding the Emergent Gameplay and Epistemic Emergence hypotheses:

1. The specific points selected along the complexity continuum overshot or undershot the mark in this particular experiment. This is plausible given that they were tweaked by hand during pre-tests based on subjective assessments of what seems "too simple", "too complex", or "emergent" based on a small
number of people. However, they are also quite extreme values at either end of the spectrum of simplicity and complexity, which makes this argument most plausible for the middle "emergent" condition. In other words, when aiming at the hypothesised "edge of chaos", I may have missed the mark and gone over that edge, or not gotten close enough to it (if it is a very sharp spike).

2. The way the independent variables (static and dynamic complexity) were operationalised were in some way inappropriate or flawed. Indeed, there are many measures of complexity in network theory, and many other features of chaotic systems apart from sensitivity to initial conditions. But, however network complexity was measured would not affect the actual topology in each condition. Therefore, it is difficult to maintain that scale-free topologies are better for engagement or learning (inherently learnable systems).

3. The underlying theory needs to be revised. The "edge of chaos" was not originally proposed in relation to human engagement or curiosity - that was an interpretation suggested by game designers. Although it was a promising place to start when looking for a formal definition of "emergent gameplay", perhaps it is simply barking up the wrong tree. Similarly, although scale-free networks have been found in various places in nature and regarded as interesting by network theorists, scale-free networks may, too, simply have been an initially promising idea that only leads to a dead end as far as emergent gameplay and inherently learnable systems are concerned.

6.9.2.1 Dynamic Complexity

For dynamic complexity, one could argue the "peak" at the edge of chaos does exist, but is much sharper than this study supposed, and was somewhere between the "low" and "medium" complexity conditions, and therefore the "medium" and "high" complexity conditions were actually "high" and "very high" complexity (see figure 6.14).
Figure 6.14: A possible alternative interpretation of the edge of chaos position could re-classify the complexity conditions in this study from "Low", "Medium" and "High", to "Low", "High" and "Very High", which could explain the results.

This is plausible given that the measure of dynamic complexity (Lyapunov exponent) for the "emergent" and "complex" conditions were much closer to each other than they were to the "simple" condition (figure 6.15). It is also consistent with a very literal, technical interpretation of the edge of chaos hypothesis. From a purely mathematical point of view one would expect the edge of chaos to occur somewhere around the zero mark for the Lyapunov exponent, since this is the separation point between stability and chaos in terms of sensitivity to initial conditions. The "emergent" condition in this study was not based on the pure mathematics of this concept, on the basis that this puts humans out of the loop and my investigation was primarily concerned with human cognition. Therefore, the emergent condition was tweaked to a particular state of chaos based on pilot tests and human feedback, rather than a pure mathematical interpretation of the concept of "the edge of chaos". Consequently, the "emergent" condition’s Lyapunov exponent was considerably above the theoretical ideal of zero that one might assume. It may be worthwhile in future studies to derive the settings for the emergent condition based on the literal mathematical interpretation of the edge of chaos, rather than tweaking it manually based on human feedback in pilot tests.
This explanation would be more consistent with my findings on engagement and implicit learning, but with the important exception of the explicit learning results. If the theory were correct, then it would account for the "medium" and "high" conditions, but the "low" condition should show higher explicit learning than the other two. I found no such effect. This alone brings into question the oft-repeated definition of emergence as "complex behaviour arising from simple rules", at least with respect to human cognition and learning, and at least with respect to the flocking algorithm.

If the flocking algorithm’s sensitivity to initial conditions does not explain the experience of emergence with flocking, then an accurate, human-centric account of its emergence remains elusive.

The flocking algorithm was used for its frequent mention in discussion of emergence (e.g. Adams and Dormans, 2012; Berrondo and Sandoval, 2015; De Wolf and Holvoet, 2005; Silverberg et al., 2013; Sweetser, 2008; Szabo, Teo, and Chengleput, 2014). However, it could be an "emergent phenomenon" without necessarily supporting "emergent gameplay" in the sense of promoting engagement and mastery. Perhaps it is emergent in a purely mathematical sense, or in a sense that is in some way qualitatively different to the concept of emergent gameplay. If there are indeed qualitatively different "kinds" of emergence, then that is a topic that de-
mands dedicated investigation. This study’s distinction between static and dynamic complexity may be a starting point.

It could be that the flocking algorithm is a kind of system that simply does not make for a particularly good game (at least, not on its own). There are other kinds of chaotic dynamic system that are worth investigating in future research. For example, stock-and-flow systems can be constructed to produce complex dynamic or chaotic behaviour (Sterman, 2000). Such a system would more naturally map on to a business management game, which bears a promising resemblance to popular simulation sandbox games such as SimCity³ and Rollercoaster Tycoon⁴, making it an excellent candidate for follow-up research.

6.9.2.2 Static Complexity

The inverted U-shaped graph of network connectivity was originally broadly considered in this research to have its Y axis corresponding to emergence. But it may be more appropriate to consider that Y axis as complexity. Therefore, if emergence arises at medium levels of complexity, then it does not arise at the halfway mark of that graph, but at one or both of the intermediate points (figure 6.16). This would mean that the actual emergence graph of network connectivity would likely have two peaks at each of those intermediate points, and low points at the start, middle, and end.

³SimCity is a franchise of games where the player builds and manages a city to grow it into a prosperous metropolis. Official website can be found here: https://www.ea.com/games/simcity
⁴The Rollercoaster Tycoon franchise is part of a family of “Tycoon” business simulations, in this case allowing players to build and manage their own theme park. Official website can be found here: http://www.rollercoastertycoon.com/
Figure 6.16: If the y axis of the network connectivity graph did not represent emergence, but complexity, then the highest level of complexity would be in the centre, and the medium (emergent) level of complexity would be at approximately the .33 and .66 marks along the x axis, creating two peaks for emergence. This interpretation could explain the results.

Assuming that a fully connected network (despite its high total walk count) is almost as simple as a completely unconnected network, then a scale-free network
STUDY 2: GAME COMPLEXITY & CURIOSITY

(with a total walk count in between the two) may not be an ideal middle-ground, but actually an extreme form of high complexity as far as a human learner is concerned. This interpretation would be more consistent with the findings of this study. One would expect that a complexity that is too high to result in a similar level of engagement to complexity that is too low, and to result in very little implicit or explicit learning due to its opaque, confusing complexity. In other words, total walk count may have a non-linear mapping on to the X axis of static complexity (where maximum TWC is equivalent to minimum TWC in terms of static complexity). This raises the question of how static complexity should best be conceptualised and measured to have relevance to how game systems cause engagement and learning. Measures other than TWC should be explored.

Therefore, this study may have overshot the mark and a more engaging level of complexity may involve a lower (or higher) total walk count than that used in the "emergent" condition in this study. This would imply not a single inverted U-shape as previously discussed, but a double-hump curve resulting from the use of total walk count as a measure of complexity. If total walk count does indeed merge the concepts of minimum and maximum, then it would loop back on itself and one would expect the pattern to mirror or repeat. A future study should investigate this hypothesis by use of such intermediate levels of complexity.

Alternatively, average path length could prove a useful measure of static complexity. If the three static complexity conditions are re-arranged not in terms of ascending total walk count, but in terms of ascending challenge when needing to accomplish goals in the system, then a different pattern is revealed (figure 6.17). A more highly connected system is more likely to provide a more direct path to a solution to any goals or problems that arise, and a star topology provides less direct paths but they at least conform to a consistent, memorable pattern. Whereas the scale-free network requires the most work to learn and recall possible paths to solutions. Whether this, or some other conception of static complexity would be more appropriate is very much an open question requiring future investigation.
The hypothesis that more curious individuals would be more engaged with more complex games was not supported. On the contrary, the pattern in the data was that individuals with more trait curiosity were more engaged regardless of complexity level, having no clear preference for dynamic complexity levels (conceptualised as sensitivity to initial conditions), and showing an apparent preference for less static complexity (conceptualised as total walk count). In contrast, individuals with less trait curiosity showed an apparent preference for higher static complexity (more connectivity in the network), and this was also the level of complexity rated as the most interesting by participants according to the pairwise comparisons. This may immediately appear to support the Challenging Complexity hypothesis (and therefore the theories of play as challenge, and curiosity as drive). However, it is important to closely examine the theories and measures of complexity. The manner in which static complexity was measured in this study was aimed at testing the scale-free network hypothesis of emergence, and therefore captured the overall connectivity (total walk count) of the causal network. However, another valid measure would be the average shortest path between two points in the network. This mea-
sure would capture how direct or indirect a path one would have to take to activate a gem, on average. Therefore, a measure such as shortest average path would be more appropriate to put the systems on a scale of ascending challenge (i.e. how many actions are required to activate gems). When re-arranged on this scale, the full connectivity system (which was high connectivity and therefore considered “high complexity” according to total walk count) is actually placed at the bottom of the scale as the lowest level of complexity, followed by the star topology as a medium form of complexity, and finally the scale-free network as high complexity (due to having a relatively long average shortest path). Given that average shortest path is a more appropriate static complexity measure for questions of challenge, it should be used to consider the hypotheses based on theories of challenge. Using this conception of static complexity, the results show a preference for low complexity (in terms of shortest average path) according to pairwise comparisons, particularly for individuals with low trait curiosity.

Considering static complexity in terms of the average shortest path offers some explanations of the results that support some of the theories reviewed. The preference for longer average shortest paths (less connectivity) among the highly curious could be explained in terms of the challenge of discovering causal linkages: A system with fewer causal links presents a greater challenge in finding them, and therefore engages more curious players to explore every possibility to find all of them. The overall preference for shorter average paths (more connectivity), especially among the less curious individuals, could be explained in terms of effectance motivation (Klimmt, Hartmann, and Frey, 2007): A more densely connected system has a higher chance of displaying a reaction to a random action from the player, and therefore, less curious players may be more satisfied with the immediate gratification of seeing something happen from almost every action they take. This demonstrates the importance of using appropriate conceptions of static and dynamic complexity that are compatible with the theories employed in a study.

Note that the emergent condition in the static complexity game was achieved by generating the causal network using a scale-free algorithm for each participant. Thus, the complexity (when operationalised as total walk count) was constant for the
STUDY 2: GAME COMPLEXITY & CURIOSITY

low and high complexity conditions, but could vary considerably for the emergent condition. This had the benefit of increasing the generalisability of the experiment - if I had used only one particular instance of a scale-free network, the results may have been due to a particular quirk of that topology, and not specifically the fact that it was scale-free. But it had the disadvantage of adding a source of variance in the data, which may explain the lack of clear findings.

Another advantage of generating the scale-free networks individually is that the results can be plotted with total walk count along the x axis, to look for a hint of an inverted u-shaped curve, or for signs of a climb or descent that hints at the ideal "emergent" total walk count higher up or lower down on the scale. However, inspecting such graphs revealed no such sloping in either direction, which suggests that scale-free topologies, or total walk count as a measure of complexity, may simply be inappropriate lines of inquiry for research into inherently learnable systems. This suggests future research should try a different approach to the problem.

This study varied the topology of the network without changing the total number of nodes in the network (five). This imposes an upper limit on the total number of links and also the total walk count of the network. It could be that scale-free topologies only show benefits to engagement or learning when there is a certain number of nodes in the network, either lower or (more likely) higher than in this study. In other words, there may have been insufficient resolution around the "low" and "medium" sections of the x axis on these TWC graphs, and an increase in the total number of nodes would increase that resolution. However, this interpretation is unlikely given the vertical spread found in these graphs.

The game’s causal network in this study was very simple in one particular sense: there was only one kind of link (interaction) between nodes (objects) - Most games have many different kinds of interactions that can occur between the different kinds of objects. The number of kinds of links in the network is an orthogonal dimension of complexity on top of that investigated in this study. Future studies should explore that dimension of structural complexity. However, it would still be worthwhile to study game systems that only have one kind of interaction, but it is
STUDY 2: GAME COMPLEXITY & CURIOSITY

qualitatively different from the interaction used in this study. The simple interaction of "activation" between the gems was selected for its abstraction, to avoid fostering preconceptions that might interfere with the learning process, but also because it does not affect the total population of objects in the game world and thereby change the probability of interactions. Another kind of interaction that is very common in games is destruction or creation (e.g. predator eating prey, or spawning offspring). This adds a layer of complexity into the system - changes to the abundance of certain objects will change the utility of different strategies, or even cause certain actions to become impossible if a certain kind of object is made "extinct" (this complication being the primary reason it was not used in this study - it could render victory impossible in uncontrolled circumstances in the implicit learning activity). Due to the ubiquity of destruction and creation interactions in games, this would be a worthwhile interaction to include in a future study, even if it is the only kind of interaction in the causal network. Another common kind of causal network is found frequently in strategy and management games: A resource dependency and conversion network. For example, a farm takes water and fertiliser to produce wheat, a windmill takes wheat and wind power to produce flour, a bakery takes flour and water to produce bread, etc. The topologies of these resource dependency networks is another worthy subject of complexity analysis and experimental manipulation.

Therefore, future research should investigate other kinds of causal network, perhaps using destruction and creation, or resource dependency and conversion, as the interaction represented in the network. Including more than one kind of interaction in the network would make the game system more realistic, and therefore this is also a promising future direction.

There has been much discussion of the the design of emergent gameplay and its benefits towards engagement (e.g. Kickmeier-rust and Albert, 2009; Sweetser, 2008). There has also been research into the difficulty students often have comprehending emergent phenomena in science (e.g. Charles and Apollonia, 2004; Chi, 2005; Chi et al., 2012; Jacobson et al., 2011; Wilensky and Novak, 2010). But before a comprehensive theory of designing emergent gameplay can be constructed, we first need to understand exactly what it is. It seems an empirically-supported definition
of emergence or emergent gameplay has yet to be established: What feature or features of a system result in emergence or emergent gameplay? The answer to this question must come before answering how to design for emergent gameplay.

One of the most commonly suggested explanations of emergence is the edge of chaos. It seems to be a particularly appealing or intuitive explanation for emergence, given its frequent discussion. However, what is intuitive is not necessarily what is correct, and this research finally put this much-discussed assumption to the test. This research was unable to find support for the edge of chaos hypothesis. The other promising explanation was found in network theory, where scale-free networks are described as some of the most interesting (e.g., Barabási, 2002; Mitchell, 2009; Wen, Kirk, and Dromey, 2007; Wen, Dromey, and Kirk, 2009). Once again, however, there was no clear support for this conception of emergence in this study. Perhaps it should come as no surprise that for emergence - a notoriously counter-intuitive concept - the most intuitively appealing explanations are insufficient. These findings do not refute information theoretic interpretations of complexity or emergence, but they do cast doubt over this conception of emergence as a phenomenon in human cognition and interpretation of stimuli.

No standardised definition of game complexity has been established that can relate curiosity to engagement. Indeed, there is currently very little data to go on. Research such as this study will provide some of the much-needed data points that will form the necessary groundwork for theories of how player curiosity is related to complexity, and thereby eventually derive formulae that can employ measures of game complexity along with measures of curiosity (like CEI-II) to calculate specific predictions of engagement and learning based on an empirically-validated theory of curiosity.

6.9.3 Limitations & Future Work

This study only tested one kind of game system for each kind of complexity. There are many kinds of systems other than a flocking algorithm that could exhibit edge of chaos dynamic complexity, just as there are many other kinds of causal network
that could exhibit a scale-free structure. Future studies should consider a similar experimental design but with different kinds of systems or genres of game to see if results differ. For example, for the scale-free network hypothesis, instead of nodes lighting up each other as a causal interaction, they could destroy or multiply each other as in an ecosystem simulation.

I measured the levels of structural and dynamic complexity in the conditions, but it is difficult to determine if they were necessarily emergent. It would be worthwhile to gather a sample of experts and laypeople to rate the level of emergence they think are present in various systems that have varying levels of complexity according to a variety of measures. This would illuminate whether certain objective measures of system properties correspond with subjective assessments of emergence.

An interesting follow-up study could more definitively test some of the hypotheses by intervening to control the factors of implicit and explicit knowledge (table 6.17). Both implicit and explicit knowledge were dependent variables in this study, but in a future study, they could alternate being dependent and independent. For instance, if learners are given upfront knowledge of the rules of the flocking algorithm at the edge of chaos, is that sufficient to give them implicit mastery? Or does it drain away all curiosity, resulting in low engagement? Additionally, when learners are given implicit mastery (e.g. through hands-on practice and maybe coaching), how does that affect explicit knowledge and engagement? Finally, one could test when they are given both forms of knowledge, or neither.
### Additional Conditions & Hypotheses

<table>
<thead>
<tr>
<th>Control Condition: Given no knowledge</th>
<th>Hypothesised Outcomes For Emergent Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implicit Knowledge: Given time to play</td>
<td>Low implicit skill score, low explicit knowledge score, high engagement</td>
</tr>
<tr>
<td>Explicit Knowledge: Given system diagram</td>
<td>High implicit skill score, low explicit knowledge score, medium engagement (less than control)</td>
</tr>
<tr>
<td>Both: Given time to play &amp; system diagram</td>
<td>High implicit skill score, high explicit knowledge score, minimal engagement</td>
</tr>
</tbody>
</table>

Table 6.17: Additional conditions and hypotheses to be explored in future work

The edge of chaos hypothesis would predict that when given nothing, there would be high inclination for voluntary play, and low implicit and explicit knowledge scores (assuming they are tested without time to explore and learn through play). When given both forms of knowledge, they should of course score high on both knowledge tests, but have a low inclination for voluntary play due to the pre-satisfaction of curiosity. When only given implicit mastery, they will have a high implicit knowledge score, low explicit score, and a medium inclination for voluntary play. When only given explicit knowledge, they would have a low implicit score, a high explicit score, and a medium inclination for voluntary play. Such a study that directly manipulated implicit and explicit knowledge could more directly test the hypothesis of the separation of these forms of learning in emergent systems. If, for example, giving learners one form of knowledge was sufficient for them to score well on the other form of knowledge, that would disconfirm the hypothesis.

Such a follow-up experiment would be able to reveal if there is an asymmetry between static and dynamic complexity. For example, we may find that with static emergence, maybe players actually are able to infer some implicit mastery purely from explicit knowledge, or vice versa.

This research therefore highlights our minimal understanding of how concepts from complexity theory relate to human cognition and engagement, specifically with respect to emergence. It also suggests some future directions for research.
to illuminate this neglected area: Other conceptions and operational definitions of complexity, other causal network topologies, and other game genres are all good places to start.

It seems for now the notoriously enigmatic concept of emergence retains its mystery. This study sought to investigate its relationship to engagement and learning to therefore understand its role in the emergent gameplay of simulation sandbox games to help in the design of inherently learnable systems for education purposes. However, emergence is a slippery concept. This study managed to turn over a few stones and search with mathematical tools, but was only able to discover where emergence was not. This lays the groundwork for a psychological science of games as systems, starting to fill in some details of what is still a vast, unwritten map. But these initial details suggest interesting new directions to explore.

### 6.10 Integrating With MMT

This study investigated whether operational definitions of complexity from complexity theory could form relevant parameters to MMT for predicting outcomes of engagement and learning. The results support this line of inquiry. Specifically, game complexity seems to function both as a play appeal factor \( (a_g) \) in formula 3.15 \((\hat{e}_p = f(a_g))\), and a learning difficulty variable \( (c) \) in formula 3.18 \((\hat{\rho}_p = l(e_p, c))\). But the hypothesised relationships between complexity and these outcomes \((f(a_g)\) and \(l(c)\)) were not supported. Despite the prominence of flow theory in the game design literature, the oft-cited concept of an ideal intermediate challenge did not prove effective in predicting the relationship between game complexity and engagement in this study (see also Klarkowski et al., 2015, and their measurements of flow - results such as these cast doubt on the concept of an ideal intermediate challenge without, of course, disproving the much larger and more complex theory of flow itself). This suggests that while complexity is a relevant factor, it either requires a slightly different operational definition, and/or a different theory for predicting how it relates to engagement and learning. Additional work is needed to develop theories to explain and predict this data – such theories would provide the functions relating play ap-
peal factors to engagement \( (f(a_g) \text{ in formula 3.15}) \), and learning difficulty with game mastery \( (l(c) \text{ in formula 3.18}) \). Experiments such as this one provide much-needed data on which to base such theories and design such functions, but additional data on different measures of curiosity and complexity, tested under different conditions, would be invaluable for the development of such theories.

This study also supported the hypothesis that trait curiosity is a relevant play appeal preference variable \( (a_p) \) that affects individual differences in how complexity affects engagement (as in formula 3.16, \( \tilde{e}_p = f(a_g, a_p) \)). But once again, more work is needed to develop the function that relates individual trait curiosity to engagement via game complexity. Although this study did not test the possibility, curiosity is also a plausible cognitive aptitude variable \( (\zeta) \) to help predict learning in formula 3.19 (\( \tilde{\rho}_p = l(e_p, \zeta, c) \)). This possible place for curiosity in MMT should be explored in future research.

### 6.11 Conclusion

This study investigated how to design inherently learnable systems, given that many people teach themselves complex games for fun, and yet have trouble understanding complex systems in the lab. The empirical experiment addressed RQ3: What is the relationship between the complexity of a game, and the player’s ability to master it? What form of complexity best arouses the curiosity of players? The literature from game design, cognitive science, and complexity theory converged to imply that emergent games may be inherently learnable systems, provoking the curiosity needed to entice players to master them (the Emergent Gameplay hypothesis). It suggested they achieve this appeal by separating implicit and explicit learning (the Epistemic Emergence hypothesis), for example with complex behaviour arising from simple rules. This was based on the edge of chaos and scale-free network hypotheses of emergence, combined with theories of curiosity as incongruity/knowledge-gap, but is also consistent with theories of play as flow, and curiosity as optimal arousal.
This study contributed empirical findings to identify a measurable form of game complexity that influences player curiosity. The Emergent Gameplay hypothesis was not supported. Nor was there support for the Challenging Complexity hypothesis that engagement would be highest for the higher level of complexity, based on theories of play as challenge, and curiosity as drive. But there was some support for the Effectance hypothesis that engagement would be highest at the lowest level of static complexity, based on the theories of juice and play as effectance. These findings required the consideration of static complexity as both the total walk count in the network, and the average shortest path, to determine the meaning of the data in relation to these diverse psychological theories. There was some support that implicit learning and engagement may co-occur at the same level of complexity - but it was a different level depending on whether one considered static or dynamic complexity. This highlights the need for more work to establish what conceptions and measures of complexity are appropriate and effective for studying how different game designs affect learning and engagement.

Trait curiosity was found to be a significant factor in engagement in different complexity levels. But more curious individuals were not interested in more complexity. Rather, the findings suggest educational game designers should cater to individuals with lower trait curiosity. To do so, it suggested educational games may want to avoid scale-free structures to their causal systems (e.g. instead having either low connectivity such as a star topology, or complete connectivity), and low sensitivity to initial conditions, to maximise implicit learning and engagement. But much future work is needed to test alternative forms of complexity and different game systems to see if these findings can be replicated, and help determine the exact nature of the relationship between curiosity and complexity.

This study is an important initial step, providing much-needed data for determining which conceptions of complexity are appropriate and useful for understanding engagement and learning in complex games, and what role trait curiosity might play in that process. With sufficient similar data, work can begin on constructing a model that takes measures of game properties such as complexity and measures of player properties such as curiosity, to produce specific testable predic-
tions of engagement and learning that generalise across games and players with similar properties.

Along with the previous study, this helps demonstrate the value and usage of MMT to investigate more specific phenomena in educational games while retaining a logically coherent place in the larger picture of how educational and serious games achieve positive outcomes - a place provided by the MMT framework. This study posited that objectively quantifiable properties of a game, such as the total walk count of the causal network, and the Lyapunov exponent as a measure of sensitivity to initial conditions, could be relevant variables of complexity that affect both engagement and learning, and this conjecture was supported. However, much work remains to elaborate the various elements of MMT in terms of prescriptive and descriptive modelling procedures, conceptions and measures of veracity, models of individual play preferences and learning aptitudes, transfer to a real-world context, etc. But this body of work lays the groundwork for how this larger project can be slowly achieved in smaller increments with narrower, deeper studies, using MMT.
Chapter 7

Discussion

This thesis proposed that educational game research can be divided into smaller subprocesses that can be studied in greater depth while still retaining relevance to the larger picture of how to design effective educational games, when that research occurs within a unifying framework. This research addressed highly complex research questions, which necessitated an interdisciplinary approach that combined theories and methods from diverse fields such as cognitive science, complexity theory, and motivational psychology. The contributions of this research help demonstrate the value of interdisciplinary research on topics such as these.

7.1 Primary Research Question

In our increasingly complex, interconnected modern world, it is interesting that people have such trouble understanding complexity in a lab, and yet teach themselves complex games in their spare time. Educational and serious games have garnered a lot of attention for their potential to achieve positive outcomes and teach complex topics. However, literature reviews and meta-analyses have found the sheer diversity of studies difficult to interpret due to the variety of approaches, methods, and outcomes. They recommended a shift away from proof-of-concept studies towards narrower studies of how specific design changes affect specific outcomes such as engagement and learning. The primary research question of this research was: How
can serious games research move from proof-of-concept studies towards building up a formal model of how specific design changes affect specific outcomes such as learning and engagement?

Without a unifying framework, it is difficult to consistently deconstruct educational and serious games research into narrower sub-processes that can be studied individually, and easy for the field at large to overlook or dismiss such narrow studies of specific phenomena as lacking relevance to the broader picture of how to design or use educational games to achieve positive outcomes. Therefore, the primary contribution of this thesis is the Model-Master-Transfer framework to address this problem: To deconstruct how serious games achieve positive outcomes into more specific sub-processes, and to specify the broader relevance of narrower studies to the field at large, to allow them to be gradually assembled to construct a comprehensive quantitative picture of the larger processes.

MMT has various practical applications and implications:

1. For Educational Game Developers:

   (a) Reporting what game design procedure was used and what the outcome was will help establish what kinds of prescriptive modelling procedures produce what kinds of games.

   (b) Report how the design of specific game properties were derived from a specific theory of play/learning.

   (c) Quantify Game Properties: In the resulting game, objectively quantify properties that are purported to have a causal relationship to engagement or learning, to test if the prescriptive modelling procedure was able to produce, for example, the intended levels of challenge, veracity, or complexity.

   (d) Adjustable Game Properties: Academics can test theories of play and learning, discovering the relationships between specific game properties and outcomes such as learning and engagement. But they can only do so if developers build into their games administrative controls to adjust the
DISCUSSION

levels of different game properties. For example, there could be a configuration file that scientists or educators could edit to manipulate the degree of challenge, complexity, or agency in the game to test different theories of play.

(e) Build in Measures of Mastery and Engagement: These can either be explicit questionnaires, or subtle, invisible measures built into the game code itself that simply record natural player behaviour to derive measures of engagement (e.g. time spent playing) or mastery (e.g. how optimally they are playing, or how high they are scoring).

(f) Report Mastery and Engagement Outcomes: This helps test the success of the prescriptive modelling procedure, and can be used to derive predictions of transfer using the MMT’s equations.

2. For Educators:

(a) Rather than just comparing the game intervention to a no-game intervention, compare the game to itself with its game properties adjusted for different groups. For example, compare a group using a low-agency version of the game to a group using a high-agency version. Measure and report which levels of the different properties were used for different groups.

(b) Measure Engagement and Mastery: This is a direct test of the effectiveness of the design of the game, but is also necessary to test theories of transfer - We need to know what percentage of their game mastery is carried over to the final test.

(c) Report Transfer: Compare real-world performance to the amount of game mastery to determine how much mastery was able to transfer to the real world.

(d) When trying to maximise educational efficiency, one can use MMT to catch early any interventions that are not likely to succeed, to spend more time on interventions more likely to succeed. For example:

i. Quantify the properties of the game that correspond to play appeal factors critical to your theories of engagement and learning (such as
Kolmogorov complexity, or degree of challenge). If the game design process failed to produce a game with the intended properties, then this can be caught early and you can try again with a different prescriptive modelling procedure, rather than going through the whole process of testing the game and finding it did not incite engagement.  

ii. Measuring the game’s veracity (derived from a theory of learning) can indicate an upper limit on the usefulness of the game: If veracity is high, then it is possible to have large transfer benefits. If veracity is low or zero, then (according MMT’s equations) that means anything learned in the game is unlikely to have any bearing on the real-world domain you are trying to teach. Therefore, a different game should be used.  

iii. Measure players’ degree of mastery of the game. If most students are only achieving low levels of mastery of the game, then (according to the equations of MMT) there is likely to be little or no improvement when it comes to transfer that mastery to the real-world domain. Therefore, the intervention can be ended right there and you can try again with a different approach, without necessarily having to actually do the transfer tests. Catching failed interventions early like this can save time and thereby increase the total number of different interventions that can be trialled.  

3. For Researchers:  

   (a) Narrow the focus of your study to investigate just one or two processes at a time with greater depth and specificity. For example, just investigating transfer, or just investigating mastery.  

   (b) Narrow the focus to the relationship between several specific variables, deriving hypotheses from specific theories of play or learning. For example, testing the theory that increasing complexity increases challenge and therefore engagement, but decreases mastery.  

   (c) Use or develop games that allow you to manipulate these variables.
(d) Use or develop objective measures for these variables, based on the theories of play or learning used in the study (e.g. quantifying agency in a game as the number of options available to the player at a time).

(e) Where possible, normalise variables to put them on an absolute scale to enable inter-study comparison and integration into MMT to build up a larger causal model. For example, compare the time participants spend playing with the amount of play time required to master the game, thereby normalising the engagement variable.

(f) If investigating the first process - modelling the domain as a game - then select or develop a specific prescriptive modelling procedure (a game design framework), follow it, and report on the outcome in terms of the properties of the resulting game, and in terms of the repeatability of the procedure (e.g. does a different team following the same procedure produce a similar game?).

(g) If investigating the final process - transfer - then use or develop a theory of learning and transfer to derive hypotheses of how game veracity and players’ mastery of the game will affect real-world performance.

It would be unrealistic for a singular study to attempt all of the above steps across modelling, mastery, and transfer. But the value of MMT is to facilitate narrower studies that each examine one or two of the above parts of the process, while maintaining relevance to the larger picture. Therefore, MMT was then applied in two studies to address two sub-problems, totalling three experiments, to demonstrate its value and the manner in which it can be used to design narrower studies of more specific phenomena while retaining relevance to the larger field. These studies focused on the process of mastering the game, exploring ways to populate the related formulas (3.15, 3.16, and 3.18) with variables such as play appeal factors ($a_g$), play appeal preferences ($a_p$), and learning difficulty variables ($c$), and functions relating those variables to outcomes of engagement and learning ($f(a_g)$ and $l(c)$). This body of work has hopefully shown that MMT is a valuable framework to identify objectively quantifiable properties of game systems to narrow the focus and test specific hypotheses. The development of methods to objectively quantify properties of
games and their features is another valuable contribution of this research - to help to establish a connection between specific game features and outcomes of engagement or learning. The first of these methods contributed was detailing a way to calculate the Kolmogorov complexity of a game.

7.2 Dynamic Causal Nets & Kolmogorov Complexity

Methods to objectively quantify game properties and features are critical to determine if changing game property X by Y amount results in a proportional change to the measured outcomes of engagement or learning. This is vital to establishing connections between specific game properties and specific outcomes, in particular by using such variables in mathematical frameworks such as those of MMT.

Therefore, another contribution of this research was to elaborate MMT in terms of how to model games in order to objectively quantify their properties, proposing Dynamic Causal Nets (DCNs) as a means to derive a measure of Kolmogorov complexity. Other modelling methods could be used to calculate complexity, or other properties other than complexity could be calculated from the game model, but this work showed one way to do so. These methods could be applied directly in future work, or serve as a guide for how to develop other tools for similar purposes or to measure other properties or features.

7.3 Research Question 2: Motivated to Lose

MMT was first applied to RQ2: How can different combinations of game features interact to cause players to engage in off-task behaviour instead of trying to win the game (and thereby likely miss the purpose of an educational or serious game)? One way this was hypothesised to happen was by a flaw in the game design resulting in conflicting play motivations, tugging the player toward different actions. This line of inquiry led to two distinct contributions: The development of the Dynamic
Probability Response (DPR) model of challenge, and empirical findings illuminating how different combinations of game features can interact to cause off-task behaviour.

MMT emphasises the role of quantifiable properties of games to derive predictions. Therefore, a model of challenge motivation was proposed, Dynamic Probability Response (DPR), that specified what game properties can be relevant to predicting player engagement based on theories of challenge motivation. It proved effective in quantifying challenge levels in different conditions and the empirical outcomes of engagement shows some initial support for DPR: Game states that were difficult to achieve were pursued, and game states that were difficult to avoid were avoided. Game states that are both difficult to avoid and achieve should be regarded neutrally (not affecting engagement), as should game states that are neither difficult to avoid or achieve, but this prediction has yet to be tested.

The study found support for the game value adoption and juice concepts of play motivation as well. This suggests these concepts may be worth working into full theories and, like DPR, quantitative models that specify certain measurable game properties (play appeal factors, in the MMT framework) as having a role in generating specific predictions of engagement.

The second contribution of this particular study was an empirical experiment to identify if different combinations of game features may interact to cause players to engage in off-task behaviour. The study found that some play motivations can distract from others, providing an explanation for the phenomenon of when players choose to lose: Challenge can distract from win-marking, and juice can distract from the combination of challenge and win-marking. For example, accidentally making the lose condition more difficult to achieve than the intended win condition could result in off-task behaviour in players. As could making failure result in a spectacular event such as an explosion or building collapse. These are ways in which educational game designers should be wary of inadvertently creating conflicting play motivations in their design. But these results did not support the hierarchy of play motivations hypothesised. More work is needed on theories of
play motivation interaction. Little is currently known about which play motivations
may interact, amplify, distract, or cancel each other out.

This study proposed and demonstrated an operational definition (DPR) of
a play appeal factor \((a_g)\) for inserting into MMT formula 3.15. But DPR does not
specify any complex function relating game difficulty to engagement. In contrast,
flow theory, for example, specifies a function relating game difficulty to engagement,
but doesn’t specify a way to measure game difficulty. Therefore, MMT enables oper-
ational definitions of play appeal factors such as DPR to be integrated with theories
of play, where one (DPR) provides the means to measure a variable \((a_g)\), and the
other (flow) provides a function to relate that variable to outcomes of engagement
\((f(a_g))\).

This study also discovered that it is indeed possible for multiple play ap-
peal factors to interact in a non-linear fashion to interfere and distract from each
other. This highlights the need for theories of how they interact, to develop functions
that can take multiple play appeal factors and generate predictions of engagement
\((f(a_g))\).

These findings on engagement can be directly inserted into MMT to elabo-
rate the process of Mastery. Improved engagement should result in improved mas-
tyery of the game and, assuming high veracity of the game, consequently improved
transfer to the real world. Thus, the next logical step was to investigate if engage-
ment and mastery co-occur under the right conditions.

7.4 Research Question 3: Complexity & Curiosity

In this second study, MMT was applied to RQ3: What is the relationship between
the complexity of a game, and the player’s ability to master it? What form of complexity
best arouses the curiosity of players? The primary contribution of this study was an
empirical experiment that identified measurable forms of game complexity that in-
fluence player curiosity. The literature from diverse domains such as game design,
complexity theory, and cognition, converged on very similar hypotheses. They sug-
gested an inverted U-shaped relationship, where the most engagement would occur with an emergent system that achieved a medium level of complexity (the Emergent Gameplay hypothesis), and that this would result in the separation of implicit and explicit learning such that learning the complex behaviour would occur independently of learning the underlying rules of the game (the Epistemic Emergence hypothesis). But despite their frequent discussion and independent convergence across multiple fields of research, these two hypotheses about emergence were not supported in this study. However, there was some support for the Inherently Learnable Systems hypothesis: The most implicit learning did co-occur with the condition rated as the most interesting by players.

In the first experiment of this study, implicit learning scores for the low dynamic complexity condition were higher than the other two conditions. Less dynamic complexity was easier to learn than medium or high. On the other hand, implicit learning scores for the medium static complexity condition were lower than the other two conditions. A scale-free topology was harder to learn than either a star topology or completely connected causal network.

Because there were no significant findings regarding engagement, a second experiment was conducted which found signs of a significant interaction between curiosity and explicit measures of preference for different levels of static complexity (but in the dynamic complexity game, there were no significant pairwise comparisons). Explicit measures of preference for high static complexity (complete topology) were higher than low static complexity (star connectivity).

There were significant interactions between curiosity and explicit measures of preference for the static complexity game. Curiosity significantly correlated with explicit measures of preference for low (star topology) and medium static complexity (scale-free topology). Curiosity also significantly correlated with average explicit preference for low dynamic complexity (less chaos, more order), but several other correlations approached significance.

If one re-arranges the conditions from ascending connectivity to ascending difficulty of making things happen one can see a different pattern that is more con-
DISCUSSION

sistent the the effectance hypothesis. This highlights the importance of future work to establish exactly how complexity (static or dynamic) should be conceptualised and measured for serious games research.

Generally, people with higher trait curiosity are less affected by differences in dynamic complexity - simple or complex, they are similarly interested (and more interested than individuals with low curiosity score). Additionally, those with higher trait curiosity were also more interested in every level of static complexity than those with low trait curiosity. Therefore, it may be advisable to focus on catering to those with less trait curiosity when designing educational games.

Combining the findings of these two experiments provides some initial support for the Inherently Learnable Systems hypothesis: That the systems that support learning and those that are more engaging. Implicit learning coincided with preference for both kinds of complexity, but at opposite levels of complexity depending on the kind of complexity. For dynamic complexity, the low dynamic complexity (less chaos, more order) condition had the most implicit learning (significant) and the highest preference scores (not significant). For static complexity, the high static complexity condition (complete topology) had the most implicit learning (significantly better than medium complexity, but not low complexity) and the highest preference scores (significant compared to both other conditions). This provides initial support for the Inherently Learnable Systems hypothesis and for the idea that multiple conceptions of complexity may be necessary to properly study engagement and learning in complex games.

This study discovered that tools from complexity theory can be used to measure aspects of game complexity that form relevant variables in MMT formulas 3.15 and 3.18, serving as both a play appeal factor \((a_g)\) and a learning difficulty variable \((c)\) to help predict outcomes of engagement and learning. But the hypothesised relationship between these variables and their outcomes were not supported \((f(a_g) \text{ and } l(c))\). Despite the prominence of flow theory in the game design literature, this study did not find support for the oft-described inverted u-shape relationship between play appeal factors and engagement (of course, this doesn’t disprove the
larger, more complex theory of flow). This suggests that other operational definitions of complexity may need to be considered and/or other theories need to be developed to provide a different function relating complexity to outcomes of engagement and learning.

The study also found that trait curiosity is a relevant play appeal preference variable \( (a_p) \) for predicting individual differences in engagement in response to game complexity \( (f(a_g,a_p)) \) in MMT formula 3.16. Therefore, future work should be dedicated to developing a specific function that can relate those variables to the outcome of engagement based on a theory of curiosity. It is also plausible that curiosity may form a cognitive aptitude variable \( (\zeta) \) in MMT formula 3.19, but that was not tested in this study, and so is another interesting area for future research.

These findings provide some initial data to help elaborate the Mastery component of MMT. Given that the Emergent Gameplay and Epistemic Emergence hypotheses were not supported, the question remains open for now as to how exactly the variables of player curiosity, game complexity, and outcomes of engagement and learning are related. But data from studies such as these provide clues with which to develop explanatory and predictive theories that can be integrated into MMT. With more work such as this, a quantitative model of how educational and serious games achieve positive outcomes can gradually be constructed.

### 7.5 Practical Applications: Educational Game Design

This thesis generated findings that are relevant to practitioners in the field of education and game design.

#### 7.5.1 Objectively Quantifiable Game Metrics for Game Design

Quantifiable properties of games as mathematical systems can be used to predict outcomes such as engagement and learning. For example, the Dynamic Probability Response of system states can be used to measure challenge. Total walk count
and Lyapunov exponent can be used as measures of game complexity (but other measures of complexity should also be explored in future).

7.5.2 Play Motivation Interactions

Games can be designed poorly such that conflicting play motivations distract players from an intended win state. For example, challenge can distract from a state visually marked as valuable, and juicy feedback can distract from a state that is visually marked as valuable and is challenging.

7.5.3 Relationship Between Complexity, Learning & Engagement

Game systems with less sensitivity to initial conditions (i.e. more order and less chaos) can enable better implicit learning. Game systems with complete connectivity in their causal structure can also enable better implicit learning (in comparison to scale-free topologies) and are considered by players to be more interesting than star and scale-free topologies (Future studies will need to test other topologies). These findings support the speculations of some authors that an inherently fun system is an inherently learnable system, but contradicts the hypotheses of some authors that medium levels of complexity would be where to find that ideal degree of complexity that leads to interesting emergence. But other system structures and definitions of complexity should be explored to confirm this.

For educational game design, these findings suggest a good strategy might be to try to make it as ordered and as fully connected as possible while still achieving the educational goal (returning to the age-old question of how to decide what to include or exclude in a simulation while maintaining realism).

7.5.4 Curiosity as a Play Motivation

Trait curiosity is a significant variable in player interest in game complexity. In general, the results suggest that those with higher trait curiosity are less affected by
DISCUSSION

game complexity, and have a generally higher level of interest regardless of game complexity (not supporting the hypothesis that highly curious players would be bored by the simple conditions and more interested in the complex conditions). In contrast, those with lower trait curiosity are more sensitive to differences in game complexity. For dynamic complexity, the results suggest that less sensitivity to initial conditions was generally considered more interesting to less curious players (not significant). For static complexity, more connectivity was considered more interesting to less curious players (but less connectivity was considered more interesting to more curious players (not significant)). Highly (completely) connected game systems may be equally interesting regardless of player trait curiosity, whereas low connectivity game systems (e.g. star topology) may be more interesting to more curious players, and less interesting to less curious players. This suggests that less curious players are less keen to find rare causal connections, and more appreciative of having more interactions immediately available, which is consistent with the play motivations of competence, effectance, and juice - if they can readily and easily make something happen, less curious players find that kind of game more interesting, possibly because it is more immediately satisfying to be able to affect the game with any interaction at all. This is consistent with the findings of study one that juice is a powerful play motivator.

7.6 Theoretical Implications

7.6.1 Conceptions of Complexity in Games

The opposing trends in the findings of study two suggest a qualitative difference between static and dynamic forms of complexity. Therefore, multiple measures of complexity may be necessary to assess how different kinds of complexity can differentially affect learning and engagement.

Definitions/measures of complexity may need to be based on concepts of challenge (as shown by the re-arrangement of the conditions revealing a U-shaped curve in experiment 2.2).
7.6.2 Emergent Games: Inherently Learnable Systems?

The findings supported the conjecture by some authors that an inherently fun system is an inherently learnable system. That point where learnability and engagement overlap was hypothesised to be at an emergent degree of complexity hovering between simplicity and complexity. But learning and interest actually co-occurred at the lowest level of dynamic complexity and the highest level of static complexity tested in this research. This conflicts with the emergence hypothesis, but also highlights a neglected area in much need of future research - what conceptions of complexity are appropriate and effective for predicting engagement and/or learning in games? There needs to be much additional work establishing standardised measures of complexity for research into game learning and engagement.

7.6.3 Are Players Curious About Complexity?

No standardised definition of game complexity has been established that can relate curiosity to engagement. Indeed, there is currently very little data to go on.

At the start of this research there was scant data with which to derive hypotheses about how player trait curiosity relates to engagement or learning when dealing with different levels of complexity, save for broad implications from general curiosity research. Research such as this thesis will provide some of the much-needed data points that will form the necessary groundwork for theories of how player curiosity is related to complexity, and thereby eventually derive formulae to detail the interactions of the player component of MMT with the game component.

7.6.4 Curiosity & Play Motivation Interactions

There are also interesting implications for play motivation interactions. The findings for engagement demonstrate that some commonalities across individuals can be detected (as in study one), whereas other effects may require accounting for individual differences such as trait curiosity (as in study two). Therefore, it may be worthwhile
for research to catalogue which play motivations (such as challenge and juice) can be detected as common across individuals, and which play motivations (such as curiosity towards complexity) are too subtle and require accounting for individual trait differences. Some play motivations may be powerful and relatively widespread, whereas others may be weaker or occur only rarely. Study one demonstrated that indeed, some play motivations can overpower others in certain circumstances, but we currently lack a theory to rank the relative power of different play motivations or otherwise detail exactly when or why one might overpower another.

7.7 Methodological Implications

These three experiments demonstrate the methodological approach and value of MMT in serious games research:

1. Breaking the process of serious games research into more narrow processes that can be studied individually in depth. In this case, more deeply focusing on how the player masters the game, and how engagement relates to learning.

2. Treating games as mathematical systems with objectively quantifiable properties that can be used to derive predictions that can be tested.

Comparing experiment 2.2 to experiments 1 and 2.1, the findings suggest that explicit measures of preference (as in experiment 2.2) and implicit measures (as in experiment 1 using time spent pursuing different options) should both be used to help detect differences as they might show up with one measure but not the other. They also suggest that when comparing the same game to itself with slight changes in design, differences may be hard to detect, and therefore it may be advisable to design the experiment to use a within-subjects design to maximise its power to detect differences in preference or engagement.
7.8 Limitations & Future Work

7.8.1 Primary Research Question: MMT

MMT specifies relationships between the major variables in how educational and serious games can achieve positive outcomes. Each of these requires follow-up work to establish what are the most appropriate theoretical conceptions and operational definitions of these variables to allow them to be measured. The equations of MMT can then be empirically tested to assess the validity of the framework.

MMT could also be further elaborated in countless ways. The DPR model of challenge proposed in this research is but one example of how the modular structure of MMT allows it to be expanded with additional details and components in future work to test more specific hypotheses while still retaining a logically coherent place in the larger picture of how educational and serious games achieve positive outcomes. For example, the player model could be elaborated with more detailed models of play preferences, playstyles, and learning aptitudes.

This research focused on the second process of MMT: Mastery. Work also needs to be done proposing and testing the reliability of both descriptive and prescriptive modelling procedures, and theories of transfer, to complete the entire picture of modelling, mastery, and transfer.

The specific predictions made possible with the above mathematical expression of MMT could be tested experimentally to determine if these equations bear accurate predictions or require reworking. Furthermore, MMT (by design) only describes the entities and processes at a high, abstract level. It therefore depends on additional theoretical tools to fill in the details – for example, a theory of play to specify the exact relationship between play appeal factors and engagement, or a descriptive modelling procedure to help ascertain a measure of veracity. Consequently, MMT inherits the limitations of such tools. For example, veracity and transfer are much-debated concepts in need of further refinement (as explained above). Similarly, the current lack of any standardised notation for game modelling (Araújo &
Roque, 2009; Koster, 2005) or replicable descriptive modelling procedure also limits the usefulness of MMT, and therefore is a crucial area for future research.

Prescriptive modelling procedures (e.g. Zee, Holkenborg, and Robinson, 2012) need to be developed and tested for several outcomes: Their consistency to produce similar game designs given different design teams working with the same source material and prescriptive modelling procedure to design a game; The degree of veracity the procedure tends to achieve; The properties of the games the procedure tends to produce. For example, a prescriptive modelling procedure might be designed specifically to achieve the optimal amount of challenge in the resulting game design. In which case, the level of challenge in the resulting game should be measured and quantified to test if the procedure was successful in its aims. This knowledge can then be integrated with studies testing different levels of challenge to determine which result in the most engagement for players.

To that end, much work is needed to formalise the myriad theories of play motivation such as challenge. These play appeal factors need to be developed into more precise models that specify exactly what quantifiable properties of games are relevant to outcomes such as engagement and learning, and thereby generate specific predictions of these outcomes using measurements of play appeal factors in specific games. This requires not only theoretical work on the nature of these play appeal factors, but also empirical work to test their predictions and tweak the models to produce accurate and reliable predictions. These can then feed back into work on prescriptive modelling procedures by indicating which game properties need to be carefully designed and in what way. Such work will also feed back into theories of individual player preferences, playstyles, and learning aptitudes, that can provide more individualised detail on how engagement and mastery occur.

Finally, theories of transfer and veracity need further empirical and theoretical work, likely alongside descriptive modelling procedures to ensure the source material and the game are modelled in a rigorous and consistent way that should theoretically capture the elements of those systems that are relevant to outcomes of engagement, learning, and transfer. Then, when enough studies have taken the
care to measure and report the needed variables, they can be assembled using MMT to gradually build up a quantitative model of how educational and serious games achieve positive outcomes.

7.8.2 Research Question 2: Motivated to Lose

This study involved brief play times and there might be different results with longer-term play behaviour. Future longitudinal studies could be designed to detect a shift in player interest to see if there is a pattern in these shifts consistently towards certain kinds of gaming activities that are suboptimal or harmful to achieving the win state.

Additionally, only one game type was used, and the variety of game genres is vast. It could be the effects observed in this study are specific to the chase game genre. Theories of challenge assume that what matters are not the specific mechanics but the degree and type of challenge they generate. Nevertheless, future studies need to investigate if similar effects can be found in different game genres.

Similarly, win-marking was a heart icon, and the results may have differed had other symbolism been used. Many games use coins, cherries, diamonds, or other more complex symbols to convey this, such as damsel in distress symbolism, or a fully-fledged narrative with backstory in order to convey which condition is the win condition. Many such win-marking symbols have very different associations. For example, a heart icon often represents health or extra lives, whereas coins often represent currency for purchasing items and upgrades. Although the assumptions of the game value adoption concept define these methods as equivalent, empirical studies should test this assumption. The results of this experiment suggest further development of the game value adoption concept may be worthwhile. A salient place to start is to determine what are the minimum necessary conditions for players to adopt the values of a game. Future work should indeed investigate the boundaries of symbolism and associations to more deeply explore the concept of game value adoption.
DISCUSSION

Juice was achieved by using one combination of effects every time (with random variation). Future research could explore different techniques to create juiciness. Although the assumptions of the juice concept are that any method of achieving a spectacular reaction is equivalent, more empirical work is needed to determine if this is so. Future work on creating a model of juice is necessary to allow us to quantify just how juicy something is, and to predict how this will affect player behaviour.

This study approached the question, ‘why do players choose to lose,’ from one of many possible angles. Future studies should investigate other possible approaches to this question, such as those that involve deliberate transgression (e.g. spoilsporting and comedic transgressive play). Here, I examined a best-case scenario to see if players could be trying to win, but misinterpret the goals of the game (and to see which motivations win out when such conflict exists). However, this does not mean that this is the only or the primary factor to explain choosing to lose as it naturally occurs in the wild. Other possible causes should be investigated in future work.

I have identified three possible play motivations to study. Each of them may be present or absent for achieving a particular game state, creating an array of possible combinations – a game state may be challenging, but not juicy and with no win marking, or any combination. There are also two game states (e.g. the win state and the lose state) to consider. Therefore, another array of possible combinations of motivators exists for the second game state in the study. Any one of the possible combinations for the first game state could be compared to any possible combination for the second game state, creating another layer of possible combinations. Furthermore, the three motivators examined herein are not exhaustive, and many other possible play motivators could be included. A study could even include a third or fourth game state, further increasing the combinatorial possibilities. Therefore, it is worth noting that this study only made use of small selection of all possible motivator combinations for game states in order to test a specific set of hypotheses. Future studies could test alternate theories of play motivation interaction that require a different selection of motivator combinations from the vast set of possibilities.
7.8.3 Research Question 3: Complexity & Curiosity

This study only tested one kind of game system for each kind of complexity. There are many kinds of systems other than a flocking algorithm that could exhibit edge of chaos dynamic complexity, just as there are many other kinds of causal network that could exhibit a scale-free structure. Future studies should consider a similar experimental design but with different kinds of systems or genres of game to see if results differ. For example, for the scale-free network hypothesis, instead of nodes lighting up each other as a causal interaction, they could destroy or multiply each other as in an ecosystem simulation.

I measured the levels of static and dynamic complexity in my conditions, but it is difficult to determine if they were necessarily emergent. It would be worthwhile to gather a sample of experts and laypeople to rate the level of emergence they think are present in various systems that have varying levels of complexity according to a variety of measures. This would illuminate whether certain objective measures of system properties correspond with subjective assessments of emergence.

An interesting follow-up study could more definitively test some of the Epistemic Emergence hypotheses by intervening to control the factors of implicit and explicit knowledge. Both implicit and explicit knowledge were dependent variables in this study, but in a future study, they could alternate being dependent and independent. For instance, if learners are given upfront knowledge of the rules of the flocking algorithm at the edge of chaos, is that sufficient to give them implicit mastery? Or does it drain away all curiosity, resulting in low engagement? Additionally, when learners are given implicit mastery (e.g. through hands-on practice and maybe coaching), how does that affect explicit knowledge and engagement?

Such a study that directly manipulated implicit and explicit knowledge could more directly test the hypothesis of the separation of these forms of learning in emergent systems. If, for example, giving learners one form of knowledge for an emergent system was sufficient for them to score well on the other form of knowledge, that would disconfirm the Epistemic Emergence hypothesis.
Chapter 8

Conclusion

We are living in an increasingly interconnected, complex world, where our collective decisions can have unanticipated indirect consequences for our environment and society. Similarly, many professions are becoming more indirect and entangled with different technologies, such as remote robotic surgery. What is today a hands-on job may soon be a job of managing variables at a computer screen. However, people seem to have tremendous difficulty handling even relatively mild levels of complexity in experiments. Yet, in the context of many modern computer games, people eagerly teach themselves vastly complex systems, all without any external guidance or coercion.

It is therefore no surprise that there has been much work trying to apply games to educational ends. However, results have been somewhat mixed, with some successes and some failures. Various literature reviews and meta-analyses have found the literature difficult to interpret due to the diversity of methods and conflicting findings. It has been difficult to compare studies and build on top of them when there is no unifying framework across diverse educational game research. This will need to change if educational game research is to make any headway. The literature reviews and meta-analyses recommend that educational and serious game research shift away from proof-of-concept studies toward narrower studies of how specific design changes affect specific outcomes such as engagement and learning.
These problems can be addressed by adopting a more formal paradigm for educational game research, where games and learning are conceptualised as systems open to mathematical description and analysis. This thesis proposed that educational game research can be divided into smaller subprocesses that can be studied in greater depth while still retaining relevance to the larger picture of how to design effective educational games, when that research occurs within a unifying framework. Comparing and extending studies would be greatly assisted by use of a unifying framework of how the entities involved - the game, the player, and the real world - interact, with a focus on relating the objectively quantifiable variables to each other. The primary research question was RQ1: How can serious games research move from proof-of-concept studies towards building up a formal model of how specific design changes affect specific outcomes such as learning and engagement? Addressing this question led to the primary contribution of this research project: Development of the Model-Master-Transfer framework. MMT was developed by identifying the major entities and processes commonly assigned a major causal role by existing frameworks collected and analysed in the literature review above. It provides a formal, mathematical framework with which to deconstruct serious games research into narrower sub-processes to be studied individually, and by which such narrow studies can be integrated into a larger mathematical structure to inform the bigger picture.

This research then demonstrated the methods and value of using MMT to address two more specific research questions in narrower studies:

2. How can different combinations of game features interact to cause players to engage in off-task behaviour instead of trying to win the game (and thereby likely miss the purpose of an educational or serious game)?

3. What is the relationship between the complexity of a game, and the player’s ability to master it? What form of complexity best arouses the curiosity of players?
In doing so, this research developed knowledge for game designers and researchers to apply: knowledge of how to make games more efficient and interesting in terms of learnability and complexity.

This aim was achieved by narrowing the focus to investigate in more depth the process of mastering a game. To avoid complications of background knowledge confounding, and improve generalisability to different domains, the games in this research were bereft of semantic embedding (i.e. they were abstract and non-representational of real-world objects). By focussing on the quantifiable mathematical properties of game systems, this research improved its generalisability and avoided common problems of qualitative interpretation and imprecision in educational game research. This enabled investigating the question of the relationship between the objectively quantifiable properties of the game, and outcomes of engagement and learning. In some cases this necessitated developing methods to objectively quantify specific game properties for a study, resulting in additional contributions of this research, such as the Dynamic Probability Response model of challenge.

This required an interdisciplinary approach, drawing upon a large and diverse body of work from fields as distinct as game design, complexity theory, and cognitive psychology. Some of the hypotheses derived from this literature were supported, but not others. The first study (addressing RQ2) supported the conjecture that the off-task behaviour could result from a flaw in the game design causing multiple play motivations to come into conflict and pull players in different directions. There was support for the challenge, juice, and game value adoption concepts of play motivation. Challenge was formalised for the study into the Dynamic Probability Response model to provide a means to objectively quantify levels of challenge in different conditions, and the support for the other two suggests they deserve more development and study into forms that also provide a means of objectively quantifying relevant properties of games. But the hierarchical structure of play motivation interaction hypothesised in the study was not supported, highlighting how little we know about how different play motivations interact or distract from each other.
The second study (addressing RQ3) involved two experiments that together provide some initial support for the Inherently Learnable System hypothesis: That those games found engaging by players are those that support learning. But support was not found for the Emergent Gameplay or Epistemic Emergence hypotheses, suggesting more work is needed on different conceptualisations of emergence and complexity to establish which are appropriate and effective for explaining and predicting outcomes of engagement and learning based on quantifiable game or system properties.

This study also found an interaction between curiosity and interest in different levels of complexity. The results suggest that educational game designers should focus on catering to individuals with lower trait curiosity, because those with high trait curiosity tend to be more interested regardless of complexity level. The less curious individual may experience more implicit learning and engagement with systems with a completely connected causal network (as compared to a scale-free topology to the causal structure), and low sensitivity to initial conditions.

The first study also has implications for educational and serious game designers. It suggests care should be taken to ensure that different play motivations don’t conflict, but align to pull players towards the one, intended goal of the game. If losing the game is made more challenging than winning, or results in a more spectacular reaction such as a catastrophic explosion, it may lure players towards trying to lose the game instead of win, which can derail the process and completely defeat the purpose of the educational game.

MMT proved a valuable framework for narrowing the focus of these studies to investigate specific psychological phenomena in more depth. Certain game properties (based on Dynamic Probability Response and measures of static and dynamic complexity) were found to result in certain outcomes for engagement and learning. This work demonstrates the value and the process of how to apply MMT to investigate how educational and serious games achieve positive outcomes. It generated some practical applications for educational game design and interesting theoretical implications for future work. This lays the groundwork for gradually
building up a comprehensive, quantitative model of educational games. However, it is a vast topic, and much work remains to elaborate and test other elements of these processes and sub-processes. But with a unifying formal framework such as MMT, there is a much clearer path toward that goal.
Appendix A

Massively Scalable Learning

The following is an extensive elaboration of a short paper published in a conference, with some passages retained verbatim (Tornqvist, Wen, and Tichon, 2017a, © 2017 IEEE).

The concept of inherently learnable systems has arisen at various points in the course of this research project. It is worthwhile elaborating exactly where and why inherently learnable systems would be useful to develop (and by extension, where they would not be so practical). Considering these applications and contexts brings various contextual and social factors into sharp focus. Modelling these contextual factors could ultimately help expand MMT to also cover researching the beneficial outcomes of serious games when used in large populations. It is an educational research paradigm I will term Massively Scalable Learning (MSL), and it emphasises a neglected set of priorities and considerations in the serious games literature.

This analysis reveals various social and contextual factors that relate to the intended beneficial outcomes of serious games, and the mechanisms by which they may do so. These factors and mechanisms could form a new social-contextual entity of MMT. That would entail formalising these factors and mechanisms into mathematical formulae and integrating them with those of MMT described in the preceding chapters. This work of formalisation and integration is ongoing, and what fol-
allows is a (non-mathematical) description of these variables and processes that were identified by considering the paradigm of Massively Scalable Learning.

A.1 Introduction

There is no shortage of research on distinguishing different player types, preferences or personalities (e.g. Barata et al., 2014; Bateman, 2014; Drachen, Canossa, and Yannakakis, 2009; Ferguson and Olson, 2013; Hamari and Tuunanen, 2014; Nacke, Bateman, and Mandryk, 2011; Olson, 2010; Yee, 2005), or different learning styles (e.g. Cassidy, 2004. But see also Kirschner, 2017; Olsen, 2006; Riener and Willingham, 2010). Such research prioritises the learning outcomes for individuals. However, tailoring lessons and games to individuals is simply not practical on the scale of large populations.

Dealing with large populations entails a very different set of priorities. While there is plentiful research on maximising individual efficiency of learning (in terms of either learning speed, learning comprehensiveness, or learning retention over time), there is little known about how to make learning scalable. This research details a highly relevant but neglected set of priorities for serious games research, outlines the current state of knowledge and the myriad unresolved questions. Mechanisms and factors are described that determine if learning scalability is relevant and effective for a given context. Existing research is evaluated to derive some initial guidelines and principles for game design prioritising scalability.

There has been much excitement recently about the potential of games for learning, and about unguided exploration in education. But, as previously detailed in the literature review of this thesis, there have been mixed results regarding serious games and unguided exploratory learning (see Ambiguity of Educational Game Outcomes section, page 27). A common concern has been how to correctly merge the educational material with the very mechanics of the game, such that mastering play of the game is synonymous with learning the desired material. Games that achieve this aim have been called conceptually integrated (Clark et al., 2015).
To achieve conceptual integration, one must consider what games do well, and what makes them games in the first place. Instead of trying to twist games to fit a traditional educational context, one could take the opposite approach and start by asking where games fit (i.e. the contextual and social factors of play). Here, I argue that games, especially those oriented around exploratory play, are better suited to applications of scalability than individual efficiency. I term this application *Massively Scalable Learning* (MSL): Education designed to occur with minimal maintenance and management, to allow it to be equally effective for a large number of unspecified individuals in uncontrolled circumstances to maximise learning saturation in a population. I define and break down MSL into its major components and examine what is known about each component to extract some initial guidelines that can be applied by practitioners of serious game design. I explain the mechanisms and factors that determine the relevance and effectiveness of MSL in a given context, and identify the most significant knowledge gaps where future research is sorely needed in this neglected area. The principles and priorities of MSL are clearest when contrasted with those of individual efficiency.

A.2 Background: The Problem

A.2.1 Individual Differences & Individual Learning Efficiency

Individual differences are an important consideration in serious games research. However, there are certain research contexts where it can be problematic or impossible to focus on individual differences, and in some cases, it could even harm the external validity of such a study.

For a hypothetical educator seeking to employ games, they have a problem: they usually cannot choose the personalities of their students. They are likely to be stuck with an inconvenient mix of different preferences, past experiences, etc. They likely have insufficient resources to make multiple different games, tailored to accommodate the unique differences of each student. Therefore, one generally-effective game is needed.
One approach is to try to design the game with the ability to alter itself automatically during play to accommodate individual differences. Dynamic difficulty adjustment is a prime example (Alexander, Sear, and Oikonomou, 2013; Baldwin et al., 2013). However, there are practical limitations in how malleable a game can be. The ambition to achieve complete accommodation for each unique case can result in the original problem of trying to make one game per person. Another approach is to try to appeal to what individuals have in common, rather than tailor to their distinct preferences. Neither approach could produce a “perfect” game, simply due to reality’s complexity and scarcity of time and resources. One cannot expect a game to be 100% effective for 100% of individuals. Instead, one needs to consider bang for buck, and try to make a game that is mostly effective for most people, given the resources available to make it.

One form of individual difference is play motivations, behaviours, or preferences (Bateman, 2014; Hamari and Tuunanen, 2014). For example, one person might be more inclined to play for social bonding. Another person might prefer to engage in competition, and thus gravitate toward competitive games. However these categories are not sharply divided (Barata et al., 2014; Drachen, Canossa, and Yannakakis, 2009; Ferguson and Olson, 2013; Nacke, Bateman, and Mandryk, 2011; Olson, 2010; Yee, 2005). One person may engage in each form of play to varying degrees.

As mentioned above, educators are rarely able to choose their students. Often educational games are deployed for a mixed population (Frank, 2011; Gallagher and Prestwich, 2013; Kim, Park, and Baek, 2009; Malone, 1980; Squire et al., 2004) with the hope that it will be generally effective despite many kinds of individual difference such as passion for gaming, different playstyles or play motivations, experience with computers, etc. If one were to screen one’s sample to limit it to, for example, experienced gamers, it could damage the external validity of the study. Educators generally cannot or do not limit their class to hardcore gamers, and therefore research on general effectiveness can often be more relevant.
Individual differences are important variables in this domain, and research on them should emphatically continue. Such findings could prove invaluable in small scale, tutor-like scenarios, where it is more practical to consider accommodating for individual differences. But there is also a need (due to how educational games are used in the field) for research on the general effectiveness of games for wide or mixed populations. At the extreme end of this spectrum is an application which can be called Massively Scalable Learning, which will be more reliant on unguided exploration than personal tutoring.

A.2.2 Unguided Exploration in Education

As detailed previously in the literature review chapter, the capacity for exploratory play to result in learning has received experimental support (e.g. Cook, Goodman, and Schulz, 2011; Gopnik and Schulz, 2007; Jennings et al., 1979; Schulz, Standing, and Bonawitz, 2008). But there has also been counter-evidence produced (e.g. Doyle, Radzicki, and Trees, 1998; Frank, 2011; Mayer, 2004) in which unguided play results in poor learning outcomes. The details can be found in the literature review section of this thesis, under the topic of exploratory play. In summary, exploratory play may not be the most efficient method of learning certain material, and may have been applied in contexts where it was not appropriate. Which raises the question: What context is more appropriate for exploratory play to be effective?

A.3 What is Massively Scalable Learning?

Exploration in serious games should not be heralded as an educational panacea. Like any tool, one must consider its particular strengths and applications. Unguided discovery in games may or may not allow one to learn more quickly, or more comprehensively, or have greater retention of knowledge, but it is undeniably more scalable. Perhaps not more efficient for teaching a specified individual, but more efficient at teaching a large number of unspecified individuals with unknown prior knowledge in uncontrolled circumstances. If one can design an inherently inter-
MASSIVELY SCALABLE LEARNING

One can just release the creation into the world like a virus (figure A.1) and, while it probably will not reach every single person, it will disperse and spread some level of knowledge through a portion of the population without further management. This is exploratory games’ benefit in efficiency: scale. This kind of education can be called Massively Scalable Learning (MSL): Education designed to occur with minimal maintenance and management, to allow it to be equally effective for a large number of unspecified individuals in uncontrolled circumstances to maximise learning saturation.

Figure A.1: Massively Scalable Learning entails knowledge saturation by a bottom-up process analogous to virus propagation. Figure reproduced with permission from Tornqvist, Wen, and Tichon, 2017a, © 2017 IEEE.

MSL downplays individual learning efficiency to instead prioritise learning saturation. Learning saturation is the total amount of learning achieved across an entire population. Where a paradigm of individual efficiency is concerned with how efficiently the maximum amount of learning can be achieved for an individual, MSL is concerned with how much learning can be achieved across a population for a given intervention (e.g. serious game). The amount learned by any given individual may not be the best that could be achieved with traditional instruction, but in terms
of surface area (population) covered, it is most efficient. When institutions can only afford one or two interventions, and achieving widespread basic competency is more useful than creating only one or two experts, MSL could prove invaluable.

A.3.1 Applications & Uses

To understand the applications for MSL, it helps to first consider the opposite. For the public good humans want to stop the spread of dangerous knowledge, such as how to synthesize small pox or build improvised explosives. MSL is most useful for knowledge that falls on the opposite end of that spectrum. Knowledge like first aid, CPR, healthy lifestyle and nutrition information, how to de-escalate potentially violent situations, how to leave a cleaner footprint on the environment, critical thinking as a form of self-defence against the persuasive methods of charlatans and demagogues: knowledge that humans have an ethical duty to spread among everyone indiscriminately to ensure maximum saturation so that people live in a better, safer world.

MSL is most valuable when a basic level of proficiency in a domain confers a sharp increase in benefit, followed by diminishing returns (figure A.2). This includes any domain with a convex proficiency-to-benefit curve. The benefit will be a function of the frequency and magnitude with which situations occur where the proficiency can be applied. For example, knowledge of some basic steps for common and dangerous first aid emergencies can save life and limb, without necessarily needing a decade of professional doctoring experience. Having a basic toolset to check for red flags for bias or poor research in the media could combat misinformation and propaganda, even without a degree in modern history or literary analysis. MSL is most valuable when there is great value in a large number of people having some knowledge because relevant situations are common, and basic proficiency is at least moderately effective in those situations. MSL is least useful if relevant situations are rare or if a high level of proficiency is required before benefits can be seen in those situations (a domain with a concave proficiency-to-benefit curve).
MSL will never be the mainstream method for specialist technical knowledge like heart surgery, copyright law, or computer science, but it is absolutely vital for “citizen of the world” knowledge. That is the domain where MSL is not just an academic curiosity, but paramount to human flourishing.

The importance of MSL has been neglected. Despite its potentially far-reaching impact and benefits, it has not featured prominently as a research topic or a design priority. As mentioned, the current emphasis in serious game research is on tailoring experiences to maximise individual efficiency. There are occasional mentions of the importance of the MSL paradigm, such as when Granic et al. (2014) point out, “Games designed for mental health interventions can reach these populations because they can be delivered to wherever clients reside, with little cost and effort.” Or when Fanetti (2012) proposes that, “One way digital games… could potentially be useful if there were a flexible and inexpensive method a student could use at their convenience”. But these examples are sparse.

Even when serious games have been used to teach the very “citizen of the world” topics listed above, scalability has not been a prominent research consideration. For example, serious games have been designed to teach about first aid (De Urturi, Zorrilla, and Zapirain, 2011; Kelle, Klemke, and Specht, 2013), critical thinking
MASSIVELY SCALABLE LEARNING

(Dunbar et al., 2014; Halpern et al., 2012), and nutrition and health (Font, Hedvall, and Svensson, 2017). These games teach knowledge that should ideally be ubiquitous among humans. Knowledge saturation would be ideal. Yet, considerations of scalability are rarely discussed in these very papers where it is most pertinent. For example, Halpern et al. (2012) included a virtual tutor system in their game that personalises the game to better fit an individual player. This may be effective when prioritising individual efficiency, but it presents some obstacles to scalability that will be explored in the following sections.

Any given feature may increase or decrease the proficiency individuals achieve with the educational object (advancing along the x axis in figure A.2), and increase or decrease the spread of the educational object and therefore the number of people it affects. The spread (number of people reached) can be considered a multiplier of whatever benefit is achieved on the y axis for individuals who use the educational object: If it only produces a small benefit for individuals, then spreading it to many individuals will not be as impactful as if it achieves a great benefit for individuals. Once again, this depends on the proficiency-to-benefit curve of the domain being taught, which in turn affects the value of different features (e.g. a virtual tutor system) that might improve individuals’ proficiency at the cost of reducing spread, or vice versa.

Whether a given feature should be added to an educational object depends on the proficiency-to-benefit curve and how much proficiency the object already imparts with its existing features: If one has already reached a proficiency level where benefits start to plateau, then efforts should switch to improving spread to maximise benefit. But if instead one has yet to climb a steep incline on the curve, then adding a feature to increase individual proficiency should be prioritised to maximise benefit. Therefore, an optimising strategy when designing an educational object will be to climb the steep slopes by adding features that increase proficiency, and then (on reaching a plateau) consolidate benefit by adding features that increase spread. But given that many features will increase one at the expense of the other, this will be difficult to balance.
The proficiency-to-benefit curve helps determine whether MSL should be a priority, but there are many additional factors that determine the effectiveness and relevance of MSL.

A.3.2 Factors Determining MSL Effectiveness

The effectiveness of MSL in a given context can be derived from a set of determining factors (figure A.3). The contextual factors (e.g. frequency of situations where proficiency is useful) will determine whether one should prioritise MSL to maximise benefits for society. The factors relating to the design of the educational object (e.g. how easy it is to share) will determine how well MSL is achieved and therefore how much it contributes to the ultimate benefits to society. For example, the general aim of MSL – knowledge saturation – can be broken down into the more specific sub-goal of ensuring knowledge can spread at all, which can be helped by making it easy to pass and share between individuals, which can be helped by designing the game to be robust to varying external conditions.
Figure A.3: The interacting factors determining the relevance and effectiveness of MSL.

MSL can apply to any kind of educational object, such as a video or a poster. But a most promising medium to consider is a digital game for the fact that it is both self-motivating and easy to digitally duplicate and share across people’s devices around the world.

Increasing the likelihood of the game being shared is an important component of MSL. But that is a question to which entire fields have been dedicated, such
as memes and marketing. Instead, I will focus on the factors that have received little attention:

1. Removing obstacles and minimising contingencies.

2. Maximising the breadth of acceptable starting proficiency by minimising the number of states.

A.4 Implementation: Design Principles & Guidelines

A.4.1 Removing Obstacles & Minimising Contingencies

The stereotypical security system for launching a nuclear missile is a set of multiple keys given to different people that need to be inserted and turned simultaneously. This increases contingencies to minimise the probability of it being used (by accident or on purpose). Designing an item to be self-contained and self-sufficient increases the ease with which it can be distributed to users, and with which users can share and redistribute it. In short, having requirements of the external world multiplies the chances of failure. There are more things that can go wrong.

Applying this principle of MSL to multiplayer games produces some interesting implications that distinguish MSL from the concept of self-organised learning (Mitra and Dangwal, 2010). Advocates of self-organised learning may or may not be correct when they profess the learning benefits of playing and learning in a group context, but there are potential drawbacks to multiplayer game designs when it comes to scalability.

MSL justifies concentrating resources on the singleplayer experience over multiplayer. It does not mean that multiplayer is a detrimental feature. But it does mean that singleplayer is a necessary feature. The more people a game requires to provide an engaging experience, the exponentially less likely people are to have an engaging experience – more things need to align in order to enable that engaging experience to happen, making it less likely. Countless multiplayer-only games have
MASSIVELY SCALABLE LEARNING

rapidly perished due to lack of existing players, making the game unplayable for newcomers who need someone to play with.

If the ratio of value between singleplayer and multiplayer is significantly in favour of multiplayer (e.g., 1:2 or 1:5), then there is a barrier to widespread initial adoption. The appeal of social technologies is often proportional to the number of people who use them (Sterman, 2000, p. 394). A telephone then it is a useless object if only one person has one. This creates a catch 22: For a multiplayer game to be widely adopted, it first needs to be widely adopted. This initial hurdle is not easy to clear.

The only definite solution is to ensure the game has a stand-alone singleplayer experience that delivers its full value without the contingency that there is already a community of players. If the singleplayer experience is a good enough reason on its own to play the game, then one need not worry about the contingency of adoption.

A.4.1.1 Stateless Learning: Maximising the Breadth of Acceptable Starting Proficiency

Often game artificial intelligence (AI) will make use of a behaviour tree, where the agent enters different behavioural modes by navigating down a tree of more and more specific behaviours. For example, when hungry it enters “food search”, then “track prey”, then “attack”, then “eat”, then exits back out to more general behaviour when complete.

In contrast, stateless AI involves no modes and stores no knowledge about what has happened previously: It has no current “state” to speak of. For example, a flocking algorithm instructs each agent how to act in the present moment. The result is very convincing flocking behaviour, even though it just observes its current situation and reacts. This algorithm has no memory of the past, and no plans for the future. It does not require entering any modes or storing any information over time. That is why it is stateless. It has no current state, only rules.
This principle could be applied to the educational aspects of a game, such that the learning features are designed to have no states: Stateless learning. The player will transition from a state of ignorance to a state of knowledge, and the game itself can have states, but the educational features will not. The game cannot be in the wrong state to deliver relevant lessons to the player. It minimises or eliminates external contingencies. It is permanently in “learning mode” because it has no other modes.

The more states an educational object has, the more likely that it will be in a state that is confusing when a new learner happens upon it. An MSL object must be robust to the random jostling of being handled and shared among learners in the real world, so that it is not prone to enter a state that it is difficult to navigate back to a state that is useful to a new learner. For example, a software toy that has essentially only one state (no levels or stages) cannot be shifted into a poor state for learning – it might only impart a modest amount of learning, but it will do so consistently.

Any object with more than one state will have a probability of being found in any one of those states after random interaction. Each state will have a probability of transitioning to each other state, assuming random interaction. And each state will likely have a different degree of efficiency in improving the user’s proficiency, given a required starting level of proficiency. For example, an introduction state will be very efficient at imparting a small amount of proficiency to users with little to no starting proficiency, but it will impart virtually zero proficiency to users who are already experts. The ideal scenario is for a novice with zero proficiency to start using the object when it is in its “introduction” state, and for that to transition to the next state capable of imparting a bit more proficiency at just the right time that the user has learned all they can from the introduction stage. Under classroom supervision this can be guaranteed. But in the uncontrolled conditions of the real world, it is unlikely that this will proceed with clockwork precision for all users in all circumstances. There will likely be interruptions, interference and sharing between individuals that can jostle the object into the wrong state to deliver learning to the user. For example, if an expert who is bored of the final state passes it along to a
novice and forgets to reset it back to the introduction, the new user will be confused and frustrated.

The frequency distribution for each state of the object can be calculated based on the probability of all possible transitions and assuming random interactions and disturbances. For example, a particular state might be very likely to occur because it is very easy for most other states to transition to it, making it a common state one might find the object in if one randomly happens upon it in the real world. One can also assume a certain distribution of starting proficiencies in the population (presumably most people will be at or near zero proficiency with many topics, or even negative if there are common misconceptions for the topic being taught). If each state also has an associated amount of learning it will deliver based on the proficiency of the user (e.g., the introduction state is only useful to novices), then one can combine that with the distribution of starting proficiencies in the population, and the frequency distribution of all the object’s states, in order to calculate an estimated amount of learning that will occur when one distributes the educational object throughout the population. For example, a very common state might actually require a very narrow range of proficiencies to deliver any learning to the user, making it very rare that the object will actually deliver learning in uncontrolled circumstances. Due to the logic of probability, the more states the object has, and the narrower the band of assumed starting proficiency for each state, the less learning one would estimate to occur. But even this estimate of total learning must be adjusted down still further, using the “share-worthiness” of the object – An object that people do not find interesting enough to share with each other will not even reach 100% of the target population and therefore will not even achieve that estimated total amount of learning.

The stochastic model just outlined could be formally described mathematically and further elaborated with factors of how frequently the object is “revisited” or “reshared” back to people who have already seen it (thereby giving it a second chance at being in the correct state to match their proficiency), and the duration of interaction required to deliver learning per state. But such detail is not necessary to draw logical implications about how design features will affect MSL.
An automated tutorial system is a partial solution to this problem. A common design keeps the tutorial distinct from the main game, as a separate, optional activity. This is not an ideal solution. Consider the paradox of the active user: the tendency for people to not bother reading instructions, preferring to get stuck right in (Carroll and Rosson, 1987; Fu and Gray, 2004; Van Nimwegen et al., 2006). This can be very problematic in uncontrolled circumstances.

A virtual tutor attempts to learn each individual user’s strengths, weaknesses, misunderstandings and learning style to provide a tailored, optimal set of progressing lessons. Tutorials and virtual tutors certainly can be very helpful. However, they are still very dependent on external conditions, and thus they are not optimal designs for maximising scalability. The virtual tutor, in particular, is clearly more concerned with individual efficiency than scalability. They require a specific sequence of some kind as the player advances through stages. In this way, virtual tutors are comparable to adaptive tutorial systems.

Tutorials could be integrated with the main game via an adaptive tutorial system that did not require the player go through a specific tutorial sequence: It could detect information relevant to the current situation and display an onscreen hint. Once the player demonstrated enough times that they knew it, it would no longer display that hint. This further reduces contingencies (it does not rely on players selecting tutorial mode). However, it does not eliminate them, it just shifts the managerial job of instruction-organisation from the distributor (teacher) to the recipients (individual learners). If one shares the game with a friend, it will not display any hints that the previous user had already learned. Even with an option to reset the hint system, this introduces an external contingency in order to properly deliver learning. With large populations, a certain percentage will inevitably forget to reset hints when they share the game, thereby reducing the effectiveness of the game as an educational object. Ideally, one should find a way to design the game in such a way that the problem of forgetting to reset hints simply is not possible. Therefore, while plenty of techniques to maximise individual efficiency may be neutral or even helpful to scalability, the virtual tutor and adaptive tutorial are examples where op-
timising for individual efficiency can come at the expense of scalability by increasing contingencies.

The design goal would be to create inherently learnable systems, or as Squire (2008) puts it, "The idea is to develop worlds that are worth understanding". MSL requires self-teaching knowledge. There are no learning stages from novice to master. It is permanently in “learning mode” because it has no other modes. From newcomer to grand master, at no point in the process does it require that the prospective learner, the external environment, or the content is in any particular state in order to learn something from the game. If this stateless learning is achieved, then a learner cannot miss a previous lesson or get left behind. It requires no resetting or management to re-enter the learning process. It is more versatile and robust in the absence of top-down management.

Both the traditional tutorial mode and adaptive contextual tutorials are clear improvements over classroom instruction in terms of scalability. Therefore, MSL can be accomplished to varying degrees. Achieving fully stateless learning or a perfectly scalable design might turn out to be impossible. But different design trade-offs and technological advances can push the design further and further toward the goal through working on the problem. After all, a game that has occasional tutorials is better than a game with some things that are indecipherable. Using duct tape to patch a hole is not ideal, but it is better than leaving the hole unpatched.

How exactly to design games to be in a permanent learning mode (to achieve stateless learning) is a field in its infancy. As described earlier, it is likely to be informed by the cognitive science of exploratory play and unguided discovery. There are currently few definitive answers concerning what conditions or design features reliably increase the likelihood of this kind of learning, but the simulation sandbox genre might be a good place to start (Tornqvist, 2014). Most of the major components of MSL are in much need of further research.
A.5 Components of MSL

As explained above, MSL can be decomposed into multiple sub-goals, which can be subdivided further. At the broadest level, MSL can be divided into three major components.

The component of learnability is, of course, fundamental to MSL. Then there is the practical component: Can it even be acquired, accessed, or used by the population? And finally, there is the motivational component: Do people want to use it and share it with others?

These components can be considered at the group or individual level. Although MSL focuses on the large-scale effect on a population, there cannot be such an effect if the game cannot succeed on the individual level for each person in that population. For example, MSL requires that a solitary individual should ideally be able to discover, procure, and master the game all on their own. But to achieve knowledge saturation in the population at large, the whole group should also be able to duplicate and share it (and be motivated to do so). In any real-world case, they may succeed with one but not the other (for example, a game might be very motivating for individuals to play and learn, but not motivate them to share it with friends). Therefore, these three components of learnable, practical, and motivational can be subdivided into their individual and interpersonal aspects to distinguish six major components to MSL, shown in table A.1.

<table>
<thead>
<tr>
<th>Major Components of MSL</th>
<th>Individual / Intrapersonal</th>
<th>Population / Interpersonal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learnable (Obstacles)</td>
<td>By self-paced discovery (unguided, stateless) - Cognitive science</td>
<td>With others - Self-organised learning</td>
</tr>
<tr>
<td>Practical (Removing Obstacles)</td>
<td>Obvious and convenient to procure and use - HCI and UX</td>
<td>Easily moved, duplicated, and shared - HCI and UX</td>
</tr>
<tr>
<td>Motivational (Increasing Incentives)</td>
<td>Fun and exciting or intriguing - Psychology of play</td>
<td>Share-worthy, e.g. funny, interesting - Memes and viral marketing</td>
</tr>
</tbody>
</table>

Table A.1: The six major components of MSL
A.5.1 Learnable

The ideal outcome is that an unspecified individual in uncontrolled circumstances will be able to learn from the game (or other MSL object, whatever it may be). Since this process cannot rely on external guidance, or guarantee that the object will be in any pre-arranged configuration, this requires unguided, stateless learning. See the above section Unguided Exploration in Education, discussing constructivism and exploratory play. Many studies have established that this kind of learning can occur (Cook, Goodman, and Schulz, 2011; Gopnik and Schulz, 2007; Jennings et al., 1979; Schulz, Standing, and Bonawitz, 2008), but there are also many where it failed to occur (Doyle, Radzicki, and Trees, 1998; Frank, 2011; Gee, 2008; Mayer, 2004). It seems very little is known about how to ensure it does (but see Tornqvist, 2014). Future research should perhaps stop trying to demonstrate that it can occur (or comparing its individual efficiency to traditional instruction), and instead investigate the mechanisms and circumstances which systematically increase its likelihood, so that educators can design better educational games of all kinds.

A similar stream of research has focussed on self-organised learning (Mitra and Dangwal, 2010): unguided learning that is necessarily social in nature, occurring in groups of students that help each other learn. For example, players can use game forums to share discoveries of underlying rule structures and discuss the advantages and contextual uses of different strategies.

Note that self-organised learning is not strictly necessary for MSL. As long as learning on the individual scale is successful, and the game is shared, then knowledge saturation is probable. As discussed above concerning multiplayer games, having a dependency on a community of players for the game to be worthwhile can be a debilitating handicap to initial adoption. But self-organised learning can be a potentially powerful amplifier of learning if the individual learning and / or spread of the game is already assured. Self-organised learning is therefore the least important of all six components considered in this section in that it is not necessary or sufficient for MSL.

What is known about MSL learnability:
• Unguided discovery learning (alone, or in groups) can be effective for learning.

What is unknown – Gaps for researchers:

• The conditions which determine if unguided discovery will be likely and effective. Research is needed to determine what game features or other conditions increase or decrease the probability of unguided learning.

• If certain game systems are specifically suited to learning by unguided discovery.

Some initial guidelines for practitioners:

• Stateless learning: Minimise the number of states the game or its learning material can be in, so that random events while it is being handled and shared cannot put the game in the wrong state to teach the user.

A.5.2 Practical

The practical component of MSL is a matter of software engineering practices (see e.g., Dix et al., 2004; McEwan et al., 2014). On the individual level, first one must know that the game exists, then it must be as easy as possible to find, then quick and easy to acquire (e.g., download and install), and finally, it must be convenient and painless to use. Part of this entails small file sizes, universal hardware compatibility, intuitive user interface design, and marketing to ensure people are aware of its existence. The former are primarily technical problems to be solved during the project’s development, and the latter are issues of psychology. Both of which are too vast to fall within the scope of this research. Readers are advised to look into the literature on marketing, human-computer-interaction, user-interface and user-experience design. But it could be very beneficial for future research to identify the practical barriers specific to game acquisition and sharing, and how to minimise them.

The interpersonal aspect of the practical component of MSL entails making it easy to spread and share. Copy protection and Digital Rights Management (DRM)
should be avoided to ensure an MSL game can be duplicated and shared among the population. Other recommendations include minimising the overall file size of the game so that moving and duplicating it is quick, and making it platform-agnostic so that it can run on almost any hardware with almost any control input (e.g., touchscreen, controller, mouse, etc). Similarly, it will be helpful if it does not require full colour vision to understand, perfect hearing, or knowledge of any particular language.

In architecture and instruction, there is the notion of universal design: A set of principles intended to maximise the usefulness and accessibility of buildings and lessons to all people, regardless of different abilities in terms of physical or mental skills (Burgstahler, 2001; Edyburn, 2010; Harris, 2014; McGuire and Scott, 2006). In that regard, there are some interesting and relevant ideas to be found therein. However, it has a different focus from MSL. MSL has the emphasis on scalability, including maximising spread by minimising management and maintenance. Universal design has the emphasis on maximum accessibility across all populations, without necessarily concerning oneself with how much extra management, guidance, cost, learning stages, or other contingencies are introduced. It prioritises tailoring to individual differences. The goal of universal accessibility is prioritised over the practicality or logistics of maximising spread.

This is most clear in the domain of education. Universal design for instruction would advise the inclusion of more stages of progression and top-down management if that is what makes the lesson more accessible to a more diverse audience. But MSL would prompt one to consider what singular design can be simple and robust enough among diverse populations to minimise top-down management once it is deployed. MSL is about bang for buck – maximising dispersion while minimising resource consumption (Indeed, minimising resource costs is itself a means to the end of maximising dispersion).

What is known about MSL practicality:

- Basic software engineering practices can help ensure the interface is not a barrier to usage.
What is unknown – Gaps for researchers:

- Specifically for serious games, what obstacles can hinder acquisition and sharing, and how best to minimise those obstacles.

Some initial guidelines for practitioners:

- Obstacles to downloading or duplicating should be avoided, such as DRM, large file sizes, specific hardware or platform requirements, etc.

- Where possible, singleplayer should be the focus to avoid the chicken-and-egg problem of multiplayer game adoption.

A.5.3 Motivational

There are many studies into the differences in individual’s motivations for play (Barata et al., 2014; Bateman, 2014; Drachen, Canossa, and Yannakakis, 2009; Ferguson and Olson, 2013; Hamari and Tuunanen, 2014; Nacke, Bateman, and Mandryk, 2011; Olson, 2010; Yee, 2005). Researchers could develop systems similar to dynamic difficulty adjustment to try to get the game to detect and adapt to different play motivations. However, as explained in the section Individual Differences & Individual Learning Efficiency, it may not be practical to try to achieve such a vast degree of malleability. Furthermore, that approach would conflict with the aim of statelessness.

What would be more useful for MSL is establishing what kinds of play motivations are broadly effective for the greatest number of people – appealing to what people have in common, rather than what makes them different. In this respect, game design is still very much an art than a science, and more systematic research is needed. For now (and possibly ultimately), it is advisable to select a form of play that best matches the subject matter for the game, to focus on those relevant forms of play to achieve conceptual integration. This should hopefully avoid the chocolate on broccoli problem. As discussed above, exploratory play has characteristics that seem particularly well-matched for MSL applications. For a deeper discussion of this, see Tornqvist (2014).
The interpersonal aspect of the motivation component refers to the “share-worthiness” of the MSL object: Is the game likely to spread due to it being funny or interesting? These are issues best informed by research on memes and viral marketing (Botha and Reyneke, 2013; Camarero and San José, 2011; Helm, 2000; Lindgreen and Vanhamme, 2005; Wilson and Consultant, 2005). General advice from that literature is that the content be made relevant to the audience (Botha and Reyneke, 2013), result in a positive emotional reaction (Botha and Reyneke, 2013; Camarero and San José, 2011; Eckler and Bolls, 2011), be surprising (Lindgreen and Vanhamme, 2005), and that the source passing it along be trustworthy (Lindgreen and Vanhamme, 2005). However, much of this literature examined viral emails or videos. The underlying social and psychological mechanisms may not be the same as those for game sharing. Once again, more research is needed in this area.

What is known about MSL motivation:

- There are many motivations for play that can be leveraged.
- Sharing of online video and email is more likely when it is relevant to the audience, produces a positive emotional reaction, surprising, and the source is considered trustworthy.

What is unknown – Gaps for researchers:

- If certain play motivations can be established as more universal and widely-applicable than others.
- If motivations to share games are different to those for online video and email.

Some initial guidelines for practitioners:

- Conceptual integration: Ensure the play motivation is well-aligned to the learning material.
- To increase motivation to share, ensure the source is viewed as trustworthy, it is relevant to the audience, and produces positive reactions or surprise.
A.6 Conclusion

Individual differences form an important piece of the puzzle to our understanding of how serious games work and how to improve them. Individual differences are consequently a popular focus in some circles of serious game research, and fall within the *player* entity of the MMT framework. Here, they serve as a counterpoint to draw attention to a very neglected collection of factors at the other end of the spectrum, and the accompanying unresolved questions. Although there are many studies about maximising individual learning efficiency, there is little known about maximising knowledge saturation in a population.

MSL is an application for which exploratory games are almost uniquely suited, playing to their strengths. MSL is defined as education designed to occur with minimal maintenance and management, to allow it to be equally effective for a large number of unspecified individuals in uncontrolled circumstances. This research examined its constituent components and implications, giving an overview of what is known about each component. I have detailed the mechanisms and factors that determine if MSL is relevant and effective for a given context, and have identified the knowledge gaps and unresolved questions of MSL, specifying areas in much need of additional research. I derived some initial guidelines that can be applied by practitioners in the field, but this is clearly a neglected area in much need of scientific attention.

More research is needed to establish what design features are necessary or sufficient for unguided learning to take place on the individual level, and for that learning material to be shared and spread to achieve knowledge saturation. The field needs a better understanding of what makes people likely to procure and use these objects of their own initiative and to share them. This depends on what design features can help make games more self-contained – the emotional psychology of how to make them self-motivating, and the cognitive science of how to make them self-teaching without external guidance or instruction. In all of these respects, game design is still very much an art than a science. This will need to change to explore the potential for MSL to improve human flourishing by spreading critical “citizen
of the world” knowledge throughout large populations – knowledge like first aid, critical thinking, and how to de-escalate potentially violent situations.

The above non-mathematical analysis of these social and contextual factors and processes will serve as the foundation for mathematical formalisation and integration with MMT.

There has been little attention given to applications of serious games to MSL or the domains which would benefit most from it. Unfortunately, there is a surplus of questions and a dearth of answers. General public welfare demands study of the design principles and challenges involved in making serious games massively scalable.
Appendix B

Cognitive Habitat, Strategy

Ecosystem: A Predictive Model

Cognitive research has found people are sometimes adept and sometimes inept at handling complexity. Complexity is a key concept in much of cognitive science, yet the field has scarcely incorporated any of the work in complexity theory. Complexity theory may generally be too abstract to easily apply to human cognition studies. Here, the problem is addressed by considering complexity in terms of the related concept of emergence, building a model of epistemic emergence, Cognitive-Habitat Strategy-Ecosystem (CHSE), to act as an overarching framework into which different conceptions of complexity and cognition can be integrated, describing how they will then interact to produce implications and predictions for cognition in complex systems. CHSE is used to derive a definition of an emergent system based on how it would be experienced as complex. This model provides value both at the micro level, by generating specific predictions, and at the macro level, through hypothesising interactions between other cognitive theories such as cognitive load, and adaptation from failure. I detail the model’s assumptions, functionality, and possible ways to measure relevant variables.

CHSE could provide considerably more detail on the process of mastery than what has been described above the previous chapters detailing the MMT frame-
work. This will require work integrating CHSE with MMT - an ongoing task for which the below formal description of CHSE lays the groundwork.

**B.1 Introduction**

As discussed in the above literature review, cognitive studies have found conditions in which people develop surprisingly efficient strategies for handling complex situations (Kahneman and Klein, 2009; Richardson and Norgate, 2014), and others in which people persist with sub-optimal strategies (Meadows and Wright, 2008; Rouwette, Größler, and Vennix, 2004; Sterman, 2000) or fundamental misconceptions about the nature of emergent phenomena (Charles and Apollonia, 2004; Chi, 2005; Chi et al., 2012; Jacobson et al., 2011; Wilensky and Novak, 2010). Such studies rarely utilise a definition from complexity theory, possibly because most definitions from complexity theory do not attempt to provide testable predictions about human cognition (see e.g. Behringer, 2009; Bonchev and Buck, 2005; Du, 2014; Emmert-Streib, 2010; Fernandez, Maldonado, and Gershenson, 2014; Gershenson and Fernandez, 2012; Manson, 2001; McAllister, 2003; Prokopenko, 2013; Wolpert and Macready, 2004). Therefore, the field could benefit from a framework that facilitates the inclusion of theories of complexity alongside theories of cognition.

This research distinguishes between an emergent phenomenon and an emergent system by also distinguishing between *micro-complexity* (encompassing traditional operational definitions of complexity), and *macro-complexity* (entailing proposing a new operational definition). These distinctions are used to construct an overarching framework into which various theories of complexity and cognition can be integrated, and describing how they will then interact to determine how a system is experienced as emergent. It models the process of interaction between the person and the system: Conceptualising an emergent system as a cognitive habitat (for an artificial intelligence it would be a computational habitat rather than cognitive) giving rise to a diverse ecosystem of competing reasoning strategies. The model, *Cognitive-Habitat Strategy-Ecosystem* (CHSE, or simply pronounced "chase"),
is informed by specific findings in cognitive science, and CHSE is used to interpret certain findings.

CHSE provides specific testable predictions at a micro level of detail, and at a macro level provides an overarching framework that describes interactions between many disparate cognitive phenomena and theories such as causal learning, cognitive load theory, automatisation, attention, dual process theory, and adaptation in the face of failure. The model’s assumptions, functionality, and possible ways to measure relevant variables are outlined. CHSE entails a range of interesting implications and predictions for cognition and learning.

General implications for cognition are discussed along with specific implications for the mechanics of adaptation and strategy switching.

### B.2 Literature Review

All relevant literature has already been discussed in the literature review section of this thesis, under the topics of complexity, and learning in complex systems (see above literature review for details). In summary, there are various proposed definitions and characteristics of emergence (e.g. Atay, 2011; Corning, 2002; Chalmers, 2002; Johnson, 2006; De Wolf and Holvoet, 2005; Deguet, Demazeau, and Magnin, 2006; Goldstein, 1999; Jost, Bertschinger, and Olbrich, 2010; Kim, 1999), such as coherence (emergent phenomena are somewhat stable. De Wolf and Holvoet, 2005; Fromm, 2005; Moncion, Amar, and Hutzler, 2010) and radical novelty (emergent phenomena are unexpected or non-deducible from their constituent parts. Buchmann, 2001; Chalmers, 2002; De Wolf and Holvoet, 2005; Henle, 2009; Hosseinie and Mahzoon, 2011). Some have suggested emergence is not sharply distinguished from related complex phenomena, but instead may be present to a greater or lesser degree (Brodu, 2008; Chalmers, 2002; Jost, Bertschinger, and Olbrich, 2010).

Philosophers and scientists still debate (Aksentijevic and Gibson, 2012; Boschetti and Gray, 2013; Brodu, 2008; Deguet, Demazeau, and Magnin, 2006; Goldstein, 1999; Ryan, 2007) whether emergence is best considered:
A an objective phenomenon (ontological emergence), falling within the domain of complexity theory or information theory; or

B a subjective phenomenon (epistemic emergence) that depends on the properties of the observer, in which case cognitive psychology may be a better tool to study emergence.

For the study of human cognition (as in the present research project), it can be useful to consider emergence a subjective phenomenon dependent on an observer. The model proposed below can thus be categorised as an epistemic model of emergence.

B.3 CHSE Model

The CHSE model aims to describe how learning of an emergent system may occur (for example, in emergent games), which requires specifying the nature of emergence.

B.3.1 Emergent Phenomena

Efficiency of prediction is a conception of emergence that has received a limited amount of attention (Aksentijevic and Gibson, 2012; Jost, Bertschinger, and Olbrich, 2010; Prokopenko, Boschetti, and Ryan, 2008; Sambrook and Whiten, 1997; Shalizi and Moore, 2003). It broadly posits that emergence is when the behaviour of a system can be predicted more efficiently by a different algorithm than by the original algorithm that produced the behaviour itself. Bishop and Trout (2005) argue a similar case: that knowledge and theories should ultimately serve humans well in practical applications. For a formula to be of value in the pursuit for human knowledge, it must be not only accurate, but also easy to implement to solve actual problems: A formula that sacrifices a bit of accuracy to gain a lot of convenience is a worthwhile epistemic goal. A deeper exploration of this definition of emergence reveals its relevance to cognitive science.
Building from these concepts, emergence can be defined as the threshold where one reasoning strategy outperforms another in terms of bang for buck: One can get as good or better predictions by investing the same or less computational (or cognitive) resources. A reasoning strategy with a greater difference in utility could be described as more emergent. Note that this definition of emergence incorporates most of the common properties of emergence listed above, such as surprise, stability, and the possibility of discontinuity between scales, making it a relatively parsimonious explanation of the diversity of phenomena often labelled as “emergent”.

Note that, where other definitions of emergence and complexity focus on properties of the system under study, here I examine the interface that mediates interaction with and understanding of the system: Reasoning strategies. A reasoning strategy in my model is a broad category that encompasses both methods of intervention to achieve a goal, and methods of passive prediction (without intervention).

A reasoning strategy takes some information about the system (given the system is in state X), and makes a prediction (the system will be in state Y, after time t). A goal is simply a pre-determined specification of the end state Y, and an intervention is just a modification of the starting conditions: “Given the system is in state X (and the agent takes action Z), it will transform the system into state Y after time t”. In principle, one could predict the outcome Y based on the starting state X and the intervention Z, or one could work backwards using the goal outcome of Y and the starting state X to determine what intervention Z would be required. For example, the very same mental model could be used (depending on which variables are specified and which variables are deduced from those provided) to predict what the aircraft will do without intervention (e.g. glide), what it will do if the pilot banks hard left, or to deduce what the pilot should do to maintain a stable trajectory and altitude. This viewpoint blurs the distinction between a prediction from observation and a goal-directed strategy, and thus allows the easy conversion and comparison between a reasoning strategy for making predictions (without intervention), and a reasoning strategy for achieving intended goals.
Reasoning strategy is a deliberately broad term, allowing it to encompass the simplest of heuristics, and the most comprehensive of mental models. Leaving these details open to elaboration allows my model to be applied using various different theories of cognition and problem-solving. For example, Instance-Based Learning Theory (IBLT) posits that skills develop as learners associate a particular perceptual cue (e.g. falling downward) with a particular motor response (e.g. pulling up on the joystick) (Fischer, Greiff, and Funke, 2012; Geddes and Stevenson, 1997; Gonzalez and Quesada, 2003; Logan, 1988). This can be contrasted with Causal Bayes Nets (CBNs). This learning theory posits that learners discover causal relationships in systems (e.g. relative air speed and density affect the drag forces on the rudder necessary for steering the aircraft) via intervention (e.g. exploratorily trying different flight manoeuvres at different altitudes and speeds), and thereby build a mental model (Gopnik and Schulz, 2007; Gopnik and Wellman, 2012; Ross, 2013; Steyvers et al., 2003). Such a mental model can simulate what would occur in novel situations never before encountered, thus giving it an advantage over the library of experiences accumulated in IBLT. Both instance libraries and mental models demonstrate the diversity of theories which can be incorporated into my model as reasoning strategies.

Note that reasoning strategies can vary in their specificity depending on how much information they include in the states. For example, a reasoning strategy may use absolutely all the information in a system state (every variable in the system is taken into account), and specify the exact outcome (predicting precisely what every variable will be). Another strategy may discard some information about the system, or only gather a few key points of information, and make predictions about some of the outcome (e.g. only some of the variables are predicted, or those predictions specify a range of possibilities). A strategy could even take zero information into account to make a prediction – this would be a static strategy that did not adapt according to different starting conditions and therefore always produced the same prediction. Similarly, a strategy might specify the exact same intervention, regardless of the starting conditions or goal (e.g. If the pitcher throws the ball high or low, fast or slow, just swing that bat as hard as you can and hope for the best). Of course,
such static strategies are likely to have a wide margin of error in their predictions (You will not always hit the ball).

B.3.2 Measuring Degree of Emergence

This defines emergence in a relative sense. There are several broad ways to measure that relative degree of emergence: objective measures would be derived from information in the reasoning strategies, and subjective measures could be collected via empirical experiment. One objective measure of similarity between two reasoning strategies is utility difference (see below section, Calculating Utility). Other objective measures to consider are the output difference and the content difference.

Output difference answers the question: Given the same inputs, how different are the outputs? This could be calculated in terms of the Euclidean distance between the output vectors, if the outputs are multidimensional. Going through all possible inputs may be possible when the possibility space is small, or when one is only interested in the strategies’ applicability to a certain set of circumstances. Using a random sampling of possible inputs will be necessary when the possibility space is too vast.

Content difference considers what specific inputs, operations, and outputs they have in common. This measure attempts to capture how they are different to process or compute (for example, two programs may take the same amount of computation, but are composed of qualitatively different kinds of computation, involving different functions, operations and structure). When the research topic involves humans, ideally such a measure should accurately capture the extent to which humans find the thinking process to be qualitatively different (e.g. when a new strategy is not just an elaboration or slight modification of an old strategy, but seemingly completely unrelated). One way to measure content difference is to assume an information processing perspective, and treat the two strategies simply as computer programs composed of variables and functions to quantify the differences between these programs. However, this may not be an accurate reflection of human cognition. Quantifying content differences in humans’ reasoning strategies will likely
remain a thorny problem for some time, leaving output and utility difference as the more realistic options for the near future.

An objective measure of the distance or difference between reasoning strategies would be useful to compare to the subjective measure provided by participants. Subjective measures could be employed to capture the intuitive sense of emergence as surprising novelty: studies could measure participants’ ratings of degree of emergence using experimental or survey data to find the threshold point at which a reasoning strategy becomes distant enough from its neighbours to be considered “significantly emergent” by experts (or even laypeople).

Doing this comparison between subjective and objective measures may be a necessary initial step in identifying the objective measure that most accurately captures this sense of strategy difference. Different objective measures could capture different dimensions of strategy difference and therefore be more appropriate to use depending on the specific research question of a particular study.

Measures such as utility difference, output difference, and various forms of content difference could prove useful in providing an objective means of quantifying emergence if one or more of them can be found to consistently correspond to humans’ subjective impression of emergence or strategy difference. Future work will need to investigate which individual or combination of measures prove most useful for this purpose, or new measures may need to be developed.

### B.3.3 Manipulating the Conditions for Emergence

Previous definitions of emergence tend to focus on variables endogenous to the system that determine if and when emergence occurs. The CHSE definition of emergence, being epistemic, adds a whole other category of variables that determine the conditions for emergence: exogenous variables, including psychological and contextual variables. Psychological variables pertain to the user of the system, and contextual variables pertain to the situation in which that user interacts with the system. This means that a system that was once simple and non-emergent, could become
emergent merely with a change in the context of its use, rather than a change in itself. Manipulation of such conditions form some of the most obvious ways in which CHSE can be applied to actual cognitive science research, and is thus open to falsification. Relevant factors have already been specified in research on Complex Problem Solving (Frensch, 1995; Funke, 1995; Funke, 2012; Liu and Li, 2012; Quesada, Kintsch, and Gomez, 2005) and Cognitive Load Theory (Naismith and Cavalcanti, 2015; Westbrook, Kester, and Braver, 2013; Wirth, Künsting, and Leutner, 2009). As discussed in the literature review, such factors tend to include the number of inputs, transparency of the system, the time delay in the system, and other properties of the system that do not specifically employ concepts from complexity theory. CHSE provides a specific place where concepts from complexity theory can be related to cognition (discussed below as micro-complexity). CHSE predicts several specific manipulations will cause emergence.

The variables that can be manipulated to affect whether one strategy out-performs another are summarised in figure B.1. The broad structure of figure B.1 has empirical support with studies showing an indirect connection between reasoning strategy accuracy and utility. Satisficing is an example of how a highly accurate reasoning strategy may impose too high a computation cost to compete with a simpler reasoning strategy. For example, one might purchase a car according to only the two most important criteria (e.g. the fuel efficiency and the purchase price), even though other, less important factors might be relevant. The result is a decision that is not optimal, but is good enough and saves time and effort. The simpler strategy achieves slightly lesser accuracy (and therefore less benefit), but that loss of benefit is more than compensated for by imposing a much lower computation cost, resulting in the simpler strategy achieving superior overall utility than the more accurate reasoning strategy. This effect has been found in various psychology studies on satisficing (Lee, 2011; O’Hara and Payne, 1998; Radner, 2001).
A full simulation of a complete mental model of the system will, by definition, produce extremely accurate predictions, making it unlikely other reasoning strategies could provide more bang for buck. Therefore, interventions that reduce the practicality of a full mental model are likely to promote emergence. There are several endogenous manipulations (changes to the system variables) that could have this effect. The most obvious way is to increase the number of variables and their interactions in the system to the point where such a computation is beyond the user’s cognitive abilities. Such a change in the system’s structural complexity increases the computation required for a full mental model, increasing its cost and lowering its utility. Another way is to make the system highly sensitive to initial conditions, rendering full simulation several steps ahead less useful since errors are likely to accumulate. Such a change to the system’s dynamic complexity decreases the accuracy of a full mental model (as long as there is any error in observing system feedback), which decreases its benefit and therefore utility. Of course, endogenous manipulations such as these are only possible when experimenters control the nature of the system, for example, by using a game that they designed. When dealing with real-world systems, exogenous variables are the only ones open to experimental manipulation.

Time pressure can be an endogenous or exogenous variable. Certain systems will impose time restrictions by their very nature, such as when driving a car.
Time pressures can also be imposed exogenously as a contextual variable, such as when one must make a decision before one’s 10am appointment. Any form of time pressure imposes a practical upper limit on the amount of computation that can be done, and therefore, which reasoning strategies are viable. Another exogenous way to adjust the amount of available cognitive resources is to impose distractions or additional, parallel tasks that require the user’s attention.

Psychological factors (those of the user) could also be manipulated, with some effort. For example, if given plenty of practice with a particular strategy until it becomes rote, then the computation cost of that strategy has been decreased, increasing its overall utility. Alternatively, if a flawed or incomplete strategy is amended with new information to improve its accuracy, then that will also increase its utility.

And finally, the utility of reasoning strategies can be manipulated directly by the imposition of certain rewards for success and penalties for failure. For example, if any darts that land within the middle 60% of the target all achieve the same prize, then there is no point incurring a higher computation cost for any accuracy higher than 60%. In contrast, a large reward for landing within 1% of the bullseye, and no reward for any less accurate shots, will likely favour very different strategies. Similar could be said of having an extreme penalty for failure, such as death. Contrast this with a scenario where there is no penalty for failure, and a user can repeatedly try a very poor strategy over and over until it works. For example, consider a game of darts which permits players to re-use darts as many times as they like without penalty, as compared to a game of darts where players must pay a fine for every dart that does not hit the bullseye.

In summary, the utility of a strategy can be manipulated by affecting its cost and / or benefit. These can be manipulated directly, or they can be affected indirectly by manipulating the accuracy of the strategy, or the computation. Figure B.1 represents these primary elements of a strategy.
B.3.4 Calculating Utility

The first curve to consider is accuracy vs. benefit. It may be that one’s only concern is accuracy, in which case accuracy = benefit and therefore x = y. In various real-world situations the curve is likely more complicated, due to a combination of natural consequences (e.g. correctly predicting a job interviewer’s reactions to your answers can get you hired for the job), and artificial extrinsic incentives (e.g. guessing within 10% of the correct number of balls in the jar wins you a new car at the carnival). The magnitude of the benefit will, of course, need to be multiplied by the probability of receiving that benefit (e.g. a 20% chance to win $100 will be converted to an expected benefit of $20). Therefore, the accuracy vs. benefit curve is heavily situation-dependent, potentially containing many discontinuities, plateaus, and other complex nuances. In actual studies, benefit may have to be operationally defined simply using whatever variables are known. For example, benefit is often operationally defined as the monetary reward offered for different levels of performance in the game participants play in the study.

The accuracy of a reasoning strategy could be measured in various ways: For example, the Brier score which verifies the accuracy of a forecast (Brier, 1950), or in terms of the Euclidean distance between the actual outcome, and the predicted or intended outcome (e.g. how far from the centre of the bullseye your shot landed). CHSE does not mandate a particular definition of accuracy or judge one as the “correct” definition – that is beyond the scope of this research. All that matters for this discussion is to assume a satisfactory operational definition of accuracy for reasoning strategies is within grasp.

With this in mind, one must consider another other important curve: computation vs. accuracy. Computation is where much of the reviewed literature on definitions of complexity (e.g. Kolmogorov complexity, or measures such as the number of inputs, outputs, and intermediate variables in the system) can be incorporated into CHSE. In the context of CHSE, such will be referred to as forms of micro-complexity as they can describe aspects of reasoning strategies, and thus should be distinguished from the macro-complexity of the strategy ecosystem of the system.
which the strategy inhabits (Macro-complexity requires a new operational definition provided in the below section, *Emergent Systems: The Cognitive Habitat & Macro-Complexity*). Just as CHSE is open to accommodate many different cognitive theories of problem solving (e.g. IBLT, or CBNs) under the category of *reasoning strategy*, it is open to the application of a wide variety of measures from complexity theory as forms of micro-complexity within CHSE. Such micro-complexity measures are useful in CHSE to determine the computational requirements of reasoning strategies, therefore informing their cost, and ultimately their overall utility.

Any given reasoning strategy will have a certain computation requirement (micro-complexity quantified using an operational definition of complexity), and will have a certain level of accuracy that can be discovered via either simulation or empirical experiment, placing that reasoning strategy at a specific point on a computation vs. accuracy graph. The accuracy scale will obviously range from 0 to 100%, but deciding on a scale for computation is less obvious. This difficult problem can be approached by considering an absolute, normalised scale of computation, where zero represents no computation (e.g. selecting a choice at random), and one represents computation matching the original system in its complexity (i.e. a full mental simulation of the system being considered). On such a scale, strategies that are more complex than the system itself are obviously identified (exceeding a value of one, which would make them inefficient strategies that should be discarded). This scale is clearer when considering discrete units of computation such as the number of variables or factors taken into account, or the number of steps ahead in time one is planning or predicting. In which case, zero represents considering no variables at all, and one represents literally taking into account absolutely all of the variables involved, and / or fully simulating every sequence of events to the full conclusion that one is interested in predicting or planning for. For example, if there are 10 variables at play, then the X axis for computation would be divided into 10 discrete segments. If one wanted to predict three steps ahead for this hypothetical system, then the X axis would have $10 \times 3 = 30$ discrete segments. In reality, there will likely be additional complicating factors. For example, a particular strategy may involve the same number of variables as another strategy, but involve performing a much more
complicated calculation with those variables. But for the purposes of explaining this model of emergence, greater clarity can be achieved by talking in these simplified terms (see also Sambrook and Whiten, 1997).

There are many ways that computation and accuracy could relate to each other on a graph. If a reasoning strategy’s curve is such that computation = accuracy, this creates a linear curve where the accuracy of predictions always matches the percentage of the system taken into consideration by the reasoning strategy. This would mean that there is exactly one way to achieve 100% accuracy, and that is to fully simulate the entire system. More commonly the curve will be more complex, dependent on empirical results and theoretical models of human cognition. On a simple level, one may speculate that only very simple strategies are often comparable to guessing, and that very complex strategies tend to have diminishing returns on accuracy.

Note that having a curve of strategy computation vs. accuracy makes an important simplification, but also that this simplification can be justified. A virtually infinite number of possible strategies can have the exact same computation cost and different accuracies, meaning the curve would be a series of speckled columns (figure B.2).
Figure B.2: A graph of strategies is likely to contain many poor, inefficient strategies. By only keeping the best available strategies, the columns of points can be reduced to a curve.

However, most of the potentially infinite strategies in a column will be random strategies with terrible accuracy. When given the choice between two strategies of the same computational cost, but different accuracies, one would choose the strategy that provides the best accuracy. Therefore, each column can theoretically be reduced to its highest point. The result is a curve of the best that each column has to offer. Similarly, when choosing between two strategies of equal accuracy, one will choose the one with the lowest computation cost. Therefore, each row could also be reduced to its leftmost point if further simplification is necessary.

There is an additional complication of not knowing every possible strategy, and using a sub-optimal strategy in a column until a better strategy is discovered.
Therefore, each column in the scatterplot represents the known strategies so that the number of strategies considered is not infinite and keeping only the top strategy in each column is justified.

Another curve to consider is computation vs. cost. For an artificial intelligence, it could be that computation = cost. But for biological computers (e.g. humans) there are certainly other conceivable possibilities described by more complex curves. For example, some people may revel in the challenge of a complex computation. Empirical studies of human cognition will have to inform the details of this curve. However, the cost does not just capture the unpleasantness of the effort required for computation, but also other costs: Extrinsic punishments (e.g. paying a fine, or received an electric shock), or natural logical consequences (e.g. death and injury are possible costs of failure to drive a vehicle competently). Of course, the magnitude of a cost should be multiplied by its probability (e.g. exceeding the speed limit may carry an 80% risk of a $100 fine, resulting in a probability-weighted cost of $80). Also, just as with quantifying benefit, actual studies often have to use whatever operational definitions are practical, such as the amount of money a participant will lose if they play poorly in the game used in the study.

Bringing together the accuracy vs. benefit, computation vs. cost, and computation vs. accuracy curves allows one to produce the curve we ultimately seek: cost vs. benefit. This utility curve indicates what strategy gives the most benefit for the least cost. Other conceptions of utility, such as expected utility theory in economics and game theory, tend to use a slightly different meaning of the term, and consequently actually inform CHSE’s definitions of benefit and cost, thereby determining utility indirectly. In CHSE, the term utility is used simply to distinguish it from either cost or benefit, to instead refer to the difference between the benefit and the cost. An elaborated form of utility is explored in a later section, and therefore this simple form of utility is called basic utility.

Let $\beta$ be basic utility, $b(z)$ be the function describing the relation between benefit and accuracy (where $z$ is the measured accuracy), and $p(O)$ be the function
describing the relation between cost (or “penalty”) and computation (where $\mathcal{C}$ is the amount of needed computation):

$$\beta = b(z) - p(\mathcal{C})$$  \hspace{1cm} (B.1)

The formula for the basic utility of a specific strategy, $s$, would be written:

$$\beta(s) = b(z(s)) - p(\mathcal{C}(s))$$  \hspace{1cm} (B.2)

Several ways to quantify accuracy were discussed above. Similarly, the benefit associated with different degrees of accuracy can in many cases be quantified in terms of the amount of reward (e.g. financial gains). However, while the amount of computation (and the associated cost) is simple to determine for a computer, it is much trickier for a brain. There are many studies using indicators of cognitive effort (Cooper-Martin, 1994; DeLeeuw and Mayer, 2008; Naismith and Cavalcanti, 2015; Paas et al., 2010), such as the time taken to come to a decision, by simple self-report, physiological measures of stress and concentration (e.g. galvanic skin response), or by getting participants to actually decide how much they are willing to pay to avoid computing the strategy themselves (Westbrook, Kester, and Braver, 2013).

The measures selected should be as comparable as possible to allow valid comparison of the cost of computation with the benefit of accuracy. But actually performing these calculations will not be necessary to derive interesting implications and testable predications for CHSE. These foundations of the utility curve will be useful in discussing how multiple strategies can fit into an ecosystem together.

### B.3.5 Emergent Systems: The Cognitive Habitat & Macro-Complexity

An emergent system supports a diverse population of competing reasoning strategies – a cognitive habitat for a diverse strategy ecosystem. If there are one or two simple and accurate reasoning strategies that apply broadly to almost all possible
scenarios within that system, then they will come to dominate how a user understands and manipulates that system, and thus the system is not experienced as complex. On the other hand, if a user is frequently surprised by novel behaviour of the system, and has to manage frequent switching between reasoning strategies, likely also needing to modify the strategies themselves and their boundaries of applicability, then that is an emergent system because it supports a more diverse strategy ecosystem. This kind of macro-complexity (the strategy diversity supported by the cognitive habitat) should be distinguished from the more conventional definition of micro-complexity (e.g. the complexity of individual strategies, making them difficult to compute, as discussed above).

B.3.5.1 Visually Analysing the Strategy Ecosystem

A utility curve can be compressed into a one-dimensional plot by subtracting cost from benefit to come up with a total utility for each strategy, and then they can be ranked vertically in a column (figure B.3). This compressed “utility column” of strategies is easier to compare to other “utility curves” side-by-side by converting other curves into columns.

![Utility Column Diagram](image)

Figure B.3: One way to compress a cost vs. benefit curve into a column is to subtract cost from benefit to get utility, and then strategies can be ranked vertically.

This process can be repeated for different situations in the system, adding the resulting utility columns along the x axis so that one has a string of columns next to each other for easy comparison (figure B.4). This can be called a viable strategy graph.
If the possibility space of a system is small enough, every different state can be enumerated along the x axis. If the possibility space is large, a random sample of possible states can be taken or the analysis limited to a certain category of states that are relevant to a particular research question (for example, only the second opening move of a game of chess). This is useful for comparing the viable strategies for a range of situations within the one system, or across different systems. A third alternative to deal with a large possibility space, has to do with how one defines a situation.

At the most fine-grained and mathematically-precise level, a situation can be defined as a point on the phase space of a system: A unique co-ordinate that includes exactly one value for all variables in the system. However, in many systems, there may be little practical difference between many of these values (e.g. there is little difference between the aircraft being 10km above sea level, and being 10km and 3 centimetres above sea level). In such cases, it could be worthwhile to use a consistent procedure for categorising the possibility space into meaningfully-similar chunks. This would achieve a coarser-grained definition of a situation, but also a definition closer to the colloquial sense of the word. When no such lines of categorisation are self-evident, it may be necessary to simply group situations by an arbitrary increment (e.g. in cubic volumes of 1x1x1km in the aircraft example), just to make the number of situations more manageable for analysis.

A viable strategy graph allows us to clearly see the openness of the system to different strategies. For example, compare the viable strategy graphs from figure B.4 and figure B.5. Figure B.5 has a lowest-common-denominator strategy that
COGNITIVE HABITAT, STRATEGY ECOSYSTEM: A PREDICTIVE MODEL

applies across all or most situations. Such a simple and sub-optimal, but universal strategy is going to be useful when someone is unable to store multiple strategies for specialised situations, or when the extra utility gained from switching to a slightly more optimal strategy is outweighed by the cognitive cost of switching strategies (see below section, Versatility Value: The Cost of Cognitive Switching).

![Figure B.5: A viable strategy graph that is mostly full is likely to reflect a system that is dominated by one or a few broadly-applicable strategies.](image)

Users defaulting to using a universally-applicable good-enough strategy results in that singular strategy becoming dominant, and thus harming the diversity of the strategy ecosystem. In an attempt to make a system more emergent (to promote a more diverse strategy ecosystem), one might try to adjust endogenous parameters of the system to produce a viable strategy graph that is much more irregular than figure B.5, such as in figure B.6. However, if switching costs are still higher than the gained utility of strategy switching, then all this will do is push down the efficiency of the lowest common denominator strategy, rather than reduce its dominance.

![Figure B.6: Reducing the overall fullness of the viable strategy graph may not eliminate a universal dominant strategy, but just push down the efficiency of that strategy](image)

At this point it is important to address several simplifying assumptions that were made. The discussion up to this point has assumed that any points occupying the same place on the Y axis are the same strategy. This is likely to be false in many
cases where a strategy’s utility may be high in some situations, but low in others. In which case, individual strategies can be identified with a unique marker (e.g. colour, in figure B.7) and this effect will be much clearer. Note that the claim still holds that, the more fully populated the viable strategy graph, the more likely that one or more strategies will apply across all or most situations.

Figure B.7: It may be necessary to label strategies that can be identified, to determine if their utility varies across situations. Here, strategies are labelled with a colour for illustrative purposes.

This provides one reason to revisit the previous step (figures B.2 & B.3) where only the topmost and leftmost points on the utility curve were kept, and consider labelling and adding back in any strategies that show up across many situations in order to see how dominant they appear to be on the viable strategy graph. The initial process involved simplifications that can provide a rough idea of the cognitive landscape, and a revisit to re-insert commonly-appearing strategies can address some of the simplifications to investigate more specific questions that arise. The viable strategy graph’s quick and clear visual display of a system’s cognitive landscape can be useful in qualitatively analysing and discussing why one system might seem more complex than another. For more precise analyses, one must consider now the variables involved and their relationships.
B.3.5.2 Macro-Complexity: Measuring Strategy Ecosystem Diversity

I previously defined an emergent system as a system that supports a diverse population of competing strategies – a cogitative habitat for a strategy ecosystem. The field of ecosystem biodiversity provides a reasonable starting point for methods to quantify the degree of strategy diversity in a system, and thereby give a measure of macro-complexity.

Much like measuring micro-complexity, the question of measuring biodiversity has many answers that are still being debated and refined (Jost, 2007; Purvis and Hector, 2000; Whittaker, 1972), with different proposed measures capturing different elements of the concept that are considered important. One of the most obvious measures is the number of species (in this case, the number of different reasoning strategies) living in a habitat. A second measure is the amount of evenness in species populations: If the habitat is filled almost entirely with one species and only a tiny minority of other species, then it is not as diverse as a habitat that has an even number of each species. Thirdly, the difference between species must be considered. For example, if all species are very similar kinds of moss, then the habitat is not as diverse as one containing various kinds of amphibians, fungi, molluscs, mammals, birds and plant.

If one can determine which strategies tend to be used for what situation in a system, one can reduce each utility column in the viable strategy graph to that single strategy that tends to be used. This could be achieved experimentally, or it could be done theoretically to get a prediction of the empirical results by computing the utility of these strategies and thus finding which ones should dominate according to their superior utility (see below section, Versatility Value: The Cost of Cognitive Switching).

Once one has done so, this allows one to consider the actual diversity of the strategy ecosystem. If the environment is represented by an image (a 2D grid), analogous to a satellite photograph of a section of rainforest, then each pixel can be coloured to represent an individual of a species (a reasoning strategy). Individuals of the same species are the same colour. Individuals of subspecies or related species are similar colours. Individuals that are very distantly related are very different
colours. This image shows in one snapshot the number of different species, how related or distinct these species are, and how clustered or intermixed they are. For example, there might be a patch of very little variety in some regions, where one species dominates. Such a snapshot can be called a *strategy ecosystem graph* (SEG) (figure B.8).

![Strategy Ecosystem Graph](image)

**Figure B.8:** An example Strategy Ecosystem Graph. The y axis no longer represents utility (as it did in the viable strategy graph). Instead the graph space simply represents the possibility space of the system as a cognitive habitat - Each X,Y location a different situation in the system. Each pixel shade represents strategy similarity across situations in the system. Clusters of similar shades reveal a lack of diversity.

This habitat need not be two-dimensional, it could be n-dimensional: Each dimension representing a relevant variable of the habitat that affects the suitability of species. These would be the parameters of the system (i.e. the phase space).

Therefore, the colour of a pixel on a SEG is used as a unique identifier of the strategy employed in that situation, and the location of a pixel on a SEG corresponds to a situation or circumstance in the system. Differences in colour correspond to differences in strategies, and distances between pixel locations represent differences in situations or circumstances in the system. Just as a pixel contains both colour and location information, a single point or cell of a SEG contains information on the strategy used, and the situation in the system’s possibility space. Such a cell on the SEG grid can be more accurately called an *eco-cell* instead of pixel to avoid confusion with image analysis and compression techniques.

Analysing the distribution of the strategies across the habitat reveals the diversity of the strategy ecosystem. One simple way to do this is with a sequential neighbourhood similarity check: For each cell in the grid, compare all adjacent cells and measure how different this cell is to its neighbours, and weight the resulting
number based on distance between the eco-cells, distance being a measure of how
different the situations are between the two eco-cells.

This can be more intuitively understood as trying to get a measure of the
bumpiness or spikiness (as opposed to smoothness or flatness) of a topological map
of a mountainous area by calculating the slope between two points. The steepness of
the slope can be found by dividing the difference in height by the distance between
those two points. For example, let \( x \) and \( y \) be two geographical locations on a map,
\( h(x) \) and \( h(y) \) be the heights of the two locations, respectively, and let \( g(x, y) \) be the
gradient or slope between two locations.

\[
g(x, y) = \frac{|h(x) - h(y)|}{|x - y|} \tag{B.3}
\]

In this analogy, the differences in strategies provide the differences in height,
and the differences in situations across the possibility space of the system represent
distances between geographical locations. Let \( s(x) \) be the strategy employed in eco-
cell \( x \), let \( \varsigma(x) \) be the circumstances or situation of eco-cell \( x \), and let \( \mu(x, y) \) be the
uniqueness of eco-cell \( x \) in relation to eco-cell \( y \).

\[
\mu(x, y) = \frac{|s(x) - s(y)|}{|\varsigma(x) - \varsigma(y)|} \tag{B.4}
\]

However, this direct application of calculating gradient to measure the di-
versity in a SEG makes some simplifying assumptions. Rather than a simple subtrac-
tion, a complex function will likely be necessary to determine the degree of differ-
ence or similarity between either a strategy, or situation. Let \( \mu(x, y) \) be the unique-
ness of eco-cell \( x \) in relation to eco-cell \( y \), let \( \sigma(s(x), s(y)) \) be our chosen function
to calculate the difference between strategies (see above section, Measuring Degree of
Emergence), and \( d(\varsigma(x), \varsigma(y)) \) be a measure of the distance between the two situations.

\[
\mu(x, y) = \frac{\sigma(s(x), s(y))}{d(\varsigma(x), \varsigma(y))} \tag{B.5}
\]
The total value of an eco-cell’s uniqueness can be found by summing together its uniqueness relative to each other eco-cell, and then normalising that value by dividing by the maximum possible uniqueness (if uniqueness is normalised, then this just means dividing by the number of eco-cells). Let $\mu(x)$ be the uniqueness of eco-cell $x$, and $\mu(x, y)$ be the uniqueness of eco-cell $x$ in relation to eco-cell $y$.

$$\mu(x) = \frac{\sum_{i=1}^{n} \mu(x, y)}{n}$$  \hspace{1cm} (B.6)

Therefore, macro-complexity (strategy ecosystem diversity) can be calculated as the sum of the uniqueness of each eco-cell. This value can be normalised by dividing it by the maximum score: where every eco-cell is 100% unique. If maximum eco-cell uniqueness is one, then this simply means dividing the sum of uniqueness scores by the number of cells in the possibility space. Let $c$ be macro-complexity of a system, and $\mu(x)$ be the uniqueness of eco-cell $x$.

$$c = \frac{\sum_{i=1}^{n} \mu(x)}{n}$$  \hspace{1cm} (B.7)

Ways to quantify strategy difference were discussed above, which leaves methods to quantify situation difference as the next topic to consider. The easiest and most direct way to quantify situation distance $d(\varsigma(x) - \varsigma(y))$ is the Euclidean distance of the vectors in the phase space. However, this is arguably too simplistic in that it does not capture the experienced similarity between situations: Two points on the phase space may be very close to each other, but result in very different behaviour in the system, such as when there are multiple attractors for a dynamic system, or with topological mixing. Similarly, two points could be very far away in the phase space, but trivially similar and could be treated almost the exact same way, for all intents and purposes. However, one must be careful here not to create a circular definition wherein situation distance is defined as the ability to apply the same strategy equally well.
Another approach to conceptualise situation similarity is in terms of who or what is having to learn the system. While a computer may well deal directly with the parameters of the phase space, a human perceives the system through the filter of their senses, and therefore situation similarity could be measured in terms of perceptual similarity. For example, a large and significant change in one dimension of the system might result in only subtle, minor perceptible changes for the human learner (e.g. a large change in the external temperature causes a slight change in the needle of the gauge in the cockpit), whereas a comparatively miniscule change in another dimension might result in large and obvious changes for the user to observe (e.g. a slight change in the pitch or bank of the aircraft causes the entire horizon to shift noticeably).

Another relevant consideration is how likely and how quickly the system is to transition between those two states. For example, two situations might be perceptually very similar, but the only way to get from one to the other is a very slow and difficult process, so they may as well have a wall between them (e.g. changing from an altitude of 10km to 0km may be quite easy, but going from 0km to -1km may first require navigating to a geographical location below sea level). One possible measure of this is the amount of time, $t$, it takes for the system, unperturbed from any external input from a user, to naturally transition from the starting state to the other state. Of course, if the starting state is within a stable attractor to a fixed point or an orbit in the phase space, then it will never escape that without external intervention, in which case this number could be infinite. With chaotic systems, it may be unclear whether the system will ever reach the new state, or just meander infinitely everywhere else, no matter how long one observes the system being simulated (depending on the nature of the specific chaotic system). As such, this measure of situation proximity is not without its flaws.

A similar measure could consider the probability of the system transitioning between the two states, given random inputs from a user (or random external shocks to the system). The larger the shocks needed to reach the other state, the less probable (or more difficult) the transition. This measure could absorb the previous measure by including the time taken as a factor rolled into the difficulty; For
example, the difficulty of reaching state B from state A equals the size of the shock required multiplied by the time it takes to eventually reach its destination. The actual difficulty would ultimately be found by averaging this number across all (or a random sampling) of possible shocks. Or in some cases, depending on the research question, it may be appropriate to take the maximum or the minimum instead of the average.

Another measure of overall ecosystem diversity worth discussing is temporal: How the ecosystem of strategies evolves over time. This would be how much the strategy ecosystem graph changes over time. It might be very static and stable. Or it might oscillate. It could be chaotic. The relative dominance of strategies could actually remain static, while the situations to which they applied shifted dramatically. Apart from fluctuating in their level of relative dominance, strategies could go extinct entirely, and new strategies could be born. This would be a measure of macro-complexity over time, or temporo-macro-complexity.

B.4 Implications for Cognition: Mechanics of Adaptation

When computation costs decrease and/or prediction benefits increase for an alternative reasoning strategy, the new strategy will be adopted when it is more efficient than the current strategy. This will result in largely distinct reasoning occurring between the old and the new situation, with each strategy having largely independent learning improvements with practice. Such improvements could take the form of decreasing computation costs by automatisation, and/or increasing prediction accuracy by a more elaborated and nuanced mental model. The important prediction is that such improvements of practice will be independent for different reasoning strategies.

The traditional conception of learning involves adding to and elaborating one’s understanding of the system, one’s mental model growing in complexity, comprehensiveness and cohesion (Boyan and Sherry, 2011; Gopnik and Schulz, 2007; Kieras and Bovair, 1964; Kim, 2015). For example, it is common to assume that new
evidence is (or should be) interpreted in a way that maximises the cohesion of one’s existing world-view (Faust, 2008). But CHSE suggests that sometimes, the model does not increase in comprehensiveness or cohesion, but branches off, or subdivides like a cell, into multiple models that grow more distinct and specialised as they develop, not more cohesive, and that this can be advantageous.

This would help explain how mental models of a single system can often have internal contradictions. This is consistent with research on mental models in design and HCI research. This line of study has revealed that people’s mental models can end up incomplete or self-contradicting, leading to poor decisions (Doyle, Radzicki, and Trees, 1998; Gentner, 2001; Norman, 1983; Oatley, 2001).

This prediction that learning improvements will be largely independent for different strategies in an emergent system, also has implications for dual process theories of cognition. Dual process theories are a family of models of cognition that split knowledge and learning into two processes: type 1 and type 2, or implicit and explicit (Evans, 2012; Harteis et al., 2012; Kahneman and Klein, 2009; Smith and DeCoste, 2000; Sun and Zhang, 2004; Sun et al., 2007). These models attempt to explain, for example, how knowing how to ride a bike, and knowing the physics of riding a bike, are not one and the same. Or more broadly, how one could select a correct answer based on intuition, or select a correct answer based on careful reasoning. Dual-process theories are controversial, attracting various critics (Osman, 2004). For example, Hassin, Bargh, and Zimerman (2009) found that automatic decisions can also be flexible and goal-dependent. Moors and De Houwer (2006) reviewed the research on automatic processes and concluded that none of the four traits examined (efficient, unintentional, uncontrollable, and unconscious) reliably co-occurred, and thus no traits were good indicators of whether or not a process was automatic. Processes classically considered automatic often lack some or all of these traits. They concluded that rather than looking for such traits as signs of automaticity, researchers should look for and report whatever specific traits are present or absent.
Although controversial, dual-process theory is nonetheless interesting to consider in the context of CHSE. It maps almost directly on to emergence as being when simple rules give rise to complex behaviour: Learning about the simple rules (e.g. being given explicit knowledge of the system’s rules) would not be sufficient to predict the complex behaviour, or having learned the complex behaviour, it would be difficult to infer the simple underlying rules. Therefore, such emergent systems could separate explicit and implicit learning into independent processes that could, theoretically, be re-united into a single process by any of the interventions discussed previously to manipulate the degree of emergence.

As already mentioned, emergence is more likely in a system where full simulation of a complete mental model is not practical. If it was practical, then such a mental model would be very accurate in virtually all situations in that system, resulting in only one dominant strategy, and therefore no emergent situations where an alternate strategy performs better (emergence is still possible, but less likely). Such a mental model is isomorphic to the system, meaning it has structural one-to-one correspondence, which is why it is aptly called a “complete simulation”, and why it is so accurate for so many situations. Such isomorphic knowledge is a good example of explicit knowledge of the system that a participant might develop in a dual process study. For example, a user might be taught the details of the algorithms behind a flocking algorithm, but that does not mean they will be able to predict or control that flock in real time. Explicit computation of the algorithm might be perfectly accurate, but it is too slow to implement in real time. Instead, one would have to spend time learning to predict or control the flock in real time to develop those implicit skills. Similarly, if one were only given that hands-on time to learn, it is unlikely one would have also developed the explicit knowledge of the algorithm.

In other words, emergence in a system is likely to cause separation of implicit and explicit knowledge, such that learning one is independent of learning the other. This prediction is concordant with the common features of ‘surprise’ and ‘non-deducibility’ of emergent phenomena. In contrast, if a system is very simple, then it is likely that being told how it works would enable one to apply that knowledge to control the system very easily, and vice versa, hands-on learning to master
the system would enable one to articulate its simple functionality. A separation between implicit and explicit learning should arise when conditions (be they endogenous to the system, or exogenous) render an isomorphic mental model unviable. These are the same conditions for emergence. For example, as previously discussed, adjusting the sensitivity to initial conditions (without changing the rules of the system) could render a full mental model impractical and therefore cause a separation of implicit and explicit learning.

B.4.1 Strategy Switching

A person’s reasoning strategies can be arranged in a network, where each node is a reasoning strategy and each link would describe the conditions for transitioning from one strategy to another.

This would help explain some aspects of human cognition. Conceptualising these strategies as a network contrasts with a completely flat, library-like conceptualisation that one might expect of a computer. A computer might simply check the current situation and select the best strategy from the list. Humans, on the other hand, will most likely already be employing a strategy of some kind, and therefore occupy a node on the network. The default action is to execute the strategy, not to check for the best one of all possible options. They will proceed with the strategy until some salient cue signals that a transition to a new strategy is advantageous. If they are familiar with the system and have several strategies already developed, then they might simply recognise that the current conditions warrant a specific transition to another strategy. However, if they are unfamiliar with the system then this salient cue will most likely take the form of a surprising event – a failure for the strategy to yield accurate predictions. This method of waiting for a salient cue to signal the need for a new strategy, is more computationally efficient than re-assessing the relevance of all possible strategies all the time. This can be referred to as strategy inertia.

Theories of attention and automatisation (Hassin, Bargh, and Zimerman, 2009; Logan, 1988; Moors and De Houwer, 2006; Osman, 2004) could be incorpo-
rated into this space of CHSE: How automatisation reduces computation cost of a reasoning strategy, increasing strategy inertia. Even if an alternate strategy is superior, if it is never even considered due to a lack of salient cues then it cannot become dominant. Therefore, one would expect that once a strategy becomes dominant, attention to alternate strategies begins to stochastically atrophy creating a positive feedback loop reinforcing the one habitual strategy: Automatisation.

Fu and Gray (2004) found that the preferred sub-optimal strategies when using software tend to be broadly-applicable and provide perceptual feedback with every action to confirm it was correctly executed. The regular perceptual feedback has the advantage that if an action is taken and the expected result does not occur (a salient cue), then the automatised procedure can abort and one can select or create a new strategy. But in many cases, it works well enough. The broad-applicability of the preferred strategy gives it greater general utility. This creates a positive feedback loop: The strategy’s cost decreases with use, and it is used very often, because it is broadly-applicable to many situations, meaning it is automatised quickly, decreasing its cost and again increasing its overall utility (for how this dominance could be quantified, see below section, *Versatility Value: The Cost of Strategy Switching*). Thus, when a new strategy is revealed that would save a single step and a few seconds of time in a very specific case, its benefits have to compete with its costs of being an unfamiliar strategy that is much more cognitively taxing than the familiar, automatised strategy. The outcome bang is not worth the computation buck. This helps explain why sub-optimal strategies can come to dominate.

Emergent systems have the ability to deviate from this phenomenon due to macro-complexity: The cognitive environment is such that, though favourite strategies develop, frequently a novel situation will prompt (or require) the user to switch between strategies or (in the case of temporo-macro-complexity) to improvise a creative solution that was not in their repertoire of known strategies.

When a strategy fails or there is surprising data, there are several choices within the framework of CHSE:

1. Persist
A Persist – Irrelevant: Ignore or discard this data. Persist with the usual strategy.

B Persist – Outlier: Assume this data point is an outlier or an anomaly, and so the reliably-successful strategy should resume its effectiveness on the next trial. But revise the accuracy of the strategy in light of this new data (e.g. noting that its accuracy isn’t quite 99%, but adjusting it down to perhaps 95%). This treats the phenomenon as a stochastic black box, paying attention only to frequency of outcome and not cause of outcome.

2. Adapt

A Adapt – New Rules: Assume that a change in the rules has occurred. This may be the most practical (or necessary) option in cases of emergence.

   i Adapt – Change of Rules: Discard the old reasoning strategies (or just the portions that are contradicted by this data) and try to figure out new ones that match this data. This seems to focus on getting results as soon as possible, not on finding the truth, given that the user does not seem concerned with why the rules appear to have changed in the first place.

   ii Adapt – Emergence: If one assumes it is a case of emergence then the course of action is very similar, but instead of completely discarding the old strategy, one simply demarcates these current conditions as being outside the domain of application of the old strategy. Therefore, a new strategy (or a variation of the old strategy) needs to be developed to handle these special conditions. The old strategy is kept but its boundaries of application are refined (i.e. its coverage on the SEG shrinks to a smaller area, as it is discovered it is not appropriate to certain situations and a new strategy is needed).

B Adapt – Incomplete Rules: Assume that one’s current strategy is incomplete. Instead of discarding the old strategy entirely (or assuming it does not apply to these special conditions), try to expand and adjust it to be able to explain both the old and the new data, and thus to broaden its domain of applicability.
These options progress on a scale of curiosity, starting with an interest in moving right along and getting the task over with (option 1a), and ending with an interest in getting to the hidden truth and attaining the fullest understanding possible (option 2b). This is analogous to the distinction between performance orientation (being concerned with getting a high score on the test) and learning orientation (concerned with improving one’s abilities) in education (Elliot and McGregor, 2001; Litman, 2008).

When one does encounter a surprising result in an otherwise familiar system, what would/should one do? Adjust down the perceived predictive power of their model, leave one’s old model aside and start building a new one (emergent strategy), or try to expand and refine one’s old model to accommodate this anomaly? That is an open question requiring further study. But some existing research points in interesting directions.

For example, learning studies have found that the main aim of the activity affects learning outcomes (Burns and Vollmeyer, 2002; Geddes and Stevenson, 1997; Künsting, Wirth, and Paas, 2011; Wirth, Künsting, and Leutner, 2009). When participants are told that the goal is to control the system to a specific outcome, they tend to develop strategies that (while effective for that specified goal) do not transfer well to achieving other goals (option 1). In contrast, when told that the goal is to learn how the system works, they generally develop a deeper understanding that empowers them to achieve new goals in new situations (option 2b). This is similar to the distinction between performance and learning orientation.

Osman (2012) had participants learn how to adjust the variables in a digital water purification simulation to achieve a certain quality of the water. In this study the rules of the water purifier did not change, but instead they told one group that they were doing very well (regardless of how well they were actually doing), and in another condition told participants that they were doing worse than average (again, regardless of how well they were doing). When they were told they were doing well, they accepted the variability of their water quality as just random noise that can be ignored (they persisted with their strategy). But when told that they were...
doing poorly, they seemed to experiment with different strategies, suggesting that they no longer accepted the variation in their results as acceptable random noise and instead were determined to identify the cause of the variability in order to eliminate it. People’s tendency to persist or adapt could be manipulated by telling them that their margin of error was too big or just fine. In CHSE terms, this refers to adjusting the benefit vs. accuracy curve by use of a threshold that marks a certain degree of accuracy or higher as “good”, and any lower accuracy as “unacceptable”.

B.4.2 Versatility Value: The Cost of Strategy Switching

The philosophy and logic literature on meta-induction (Schurz, 2008) explores how an actor should select among many prediction methods to maximise their chances of predicting correctly. For example, how one might select among many weather forecast services, weather simulations, or strategies for forecasting the weather, to adopt the predictions that are most likely to be correct. However, such work is generally normative rather than positive, describing what an ideal rational agent ought to do, and consequently there is less focus on human elements such as the cognitive cost and inertia of switching strategies, or the emotional rewards for success or failure. CHSE considers how such psychological factors could give rise to a diverse ecosystem of competing reasoning strategies.

Just as with operational definitions from complexity theory and theories from cognitive science, CHSE does not seek to replace or supersede meta-induction. Meta-induction theories could be useful to augment CHSE in terms of discriminating the most accurate strategy, but selecting and integrating a theory of meta-induction is beyond the scope of this research, and not necessary for the present discussion. Instead, I now focus on the psychological factors that affect strategy switching.

An additional computation cost incurred in an emergent system is the cost of strategy switching: Monitoring for a salient cue or set of conditions. More complicated criteria for switching will incur a greater computational cost, and subtler perceptual cues will incur a greater attentional cost. Having a larger number of po-
tential transitions to consider also increases computational cost. That is to say, if
a node in the strategy network has many different links to other strategies to denote
possible transitions to other strategies, then keeping track of all those different sets
of transition conditions will be more taxing than a node that has only one link to one
other strategy.

This generates the interesting prediction is that a user’s strategy network
will tend towards fewer connections per node in order to minimise monitoring for
switching conditions, and therefore prefer to string strategies in series where possi-
ble, creating long chains with few branching points (figure B.9).
A flat library of strategies maximises monitoring costs for switching, because all strategies are always checked.

Having many possible transitions from one strategy to another can be useful, but too many possible transitions can impose too high a monitoring cost.

Consequently, one will try to minimise possible transitions to monitor, causing the formation of long chains in the network of strategy transitions.

Figure B.9: The cost of strategy switching suggests possible transitions to monitor should be minimised, creating long chains in a strategy transition network.
It also means that the extent to which this holds true will also depend on the relative computational cost of a given strategy in the situation. For example, a very complicated strategy will leave few resources left for strategy switching, so will prefer to have fewer linkages. On the other hand, a very simple strategy will leave plenty of cognitive resources available and so can support many more linkages to other strategies (however, people will still likely prefer to minimise cognitive effort where possible). The extent to which active monitoring is necessary, is likely inversely proportional to cue salience: If the cues that indicate the need for a transition are subtle, then one will have to pay close attention, but if the cues are jarringly obvious and attention-grabbing, then actively monitoring may not be necessary and one can simply wait for it to become obvious.

The actual act of mentally switching to another strategy is likely to incur a cost as well, due to breaking the mental routine of the previous strategy (see Moradzadeh, 2009 for a discussion of cognitive switching).

Finally, there is a possible cost in long-term memory for storing and retrieving strategies. A bigger, more complex strategy might take more cognitive resources to retrieve from memory to use. Of course, this would likely be mitigated by automation. It is more efficient to have a small number of simple strategies that can be recalled quickly and do not need frequent switching, than to have a vast number of complicated strategies that require constant switching.

The overall cost of strategy switching can be termed \textit{strategy management}, given that it is composed of two distinct costs: Monitoring for switching cues, and the actual process of switching strategies. Note that monitoring is a constant, ongoing cognitive cost, likely ameliorated by cue salience. Whereas strategy switching is a one-time cost, only incurred when the time comes to actually switch to another strategy. This cost is mitigated by the probabilistic frequency of switching – the less frequently one has to switch, the less switching cost will be an issue.

The SEG for a system can be used not just to derive a measure of the macro-complexity of the system, but to help predict individual behaviour when dealing with that system. Primarily, it helps determine strategy dominance based on strat-
egy versatility. Formula (B.1) introduced above provided a way to calculate the basic utility of a strategy based on its computation cost and accuracy, but a strategy’s versatility is also a relevant variable to its overall utility — A strategy with a wide area of applicability (even if it does not perform optimally in most of those situations where it can apply) can save computation by limiting strategy switching. Thus, there is a secondary level of computation cost and utility at play that has to be layered on top of the basic utility considered in formula (B.1). This can be called secondary utility to distinguish it from basic utility defined above. It is actually simpler to describe this level of utility not in terms of the benefits from versatility, but the costs of specialisation. The primary additional variable at this layer is the management cost of the strategy: The cognitive cost imposed by the need to administrate, monitor, manage, and switch to other strategies (the inverse of the benefits gleaned from the versatility of a strategy).

Let \( u(s) \) be the secondary utility of strategy \( s \), let \( \beta(s) \) be the basic utility of the strategy as explained above (see the above section, Calculating Utility), and \( m(s) \) be the strategy management cost of the strategy.

\[
 u(s) = \beta(s) - m(s) \tag{B.8} 
\]

As basic utility was defined above, what remains is to determine management costs based on a SEG. Total strategy management cost for a strategy could be broken down into monitoring costs and switching costs. Monitoring costs are ongoing, imposed constantly by the need to check or consider if one might need to switch strategies. Switching costs, on the other hand, are transient spikes in cost imposed by the actual act of switching from one mental routine to another. Let \( m(s) \) be total strategy management cost of a strategy \( s \), and \( \omega(s) \) be the monitoring cost of the strategy, and \( \gamma(s) \) be the switching cost of the strategy.

\[
 m(s) = \omega(s) + \gamma(s) \tag{B.9} 
\]
Each of these variables could be broken down further into more specific elements for study. For example, the monitoring cost of a strategy could be composed of the costs imposed by the subtlety of the cues one needs to be mindful of (subtler cues imposing a higher cognitive cost to monitor for), and the complexity of the criteria or calculation needed to determine if switching would be appropriate (for example, “switch if the speedometer goes higher than 60kph”, is a much simpler criterion to calculate than, “switch if the reading of the speedometer is less than half the square root of the fuel gauge, plus 53”). The cost of switching itself could also be broken down into the cost imposed by breaking the habit of the current strategy (a very familiar strategy that has been used implicitly or automatically for a long stretch of continuous time may be more difficult to stop than a very unfamiliar or esoteric strategy that required constant conscious effort to enact), and the cost imposed by the act of retrieving the new strategy (if one is switching to a very simple, intuitive, and familiar strategy, it may be easy to retrieve, but if one is switching to an entirely unfamiliar or very complex strategy, it may be considerably more difficult). Different combinations of these factors could result in different monitoring and switching costs (tables B.1 and B.2).

<table>
<thead>
<tr>
<th>Monitoring Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cue Perceptual Subtlety/Salience</td>
</tr>
<tr>
<td>Clear, Obvious Cues</td>
</tr>
<tr>
<td>Subtle, Obscure Cues</td>
</tr>
<tr>
<td>Criteria Complexity</td>
</tr>
<tr>
<td>Simple Criterion</td>
</tr>
<tr>
<td>Low cue cost. Low criteria cost.</td>
</tr>
<tr>
<td>High cue cost. High criteria cost.</td>
</tr>
<tr>
<td>Complex Calculation</td>
</tr>
<tr>
<td>Low cue cost.</td>
</tr>
<tr>
<td>High cue cost.</td>
</tr>
</tbody>
</table>

Table B.1: The micro-complexity of the criteria and the perceptual subtlety or salience of relevant cues determine the cost of monitoring for the potential need to switch strategies.
**Switching Costs**

<table>
<thead>
<tr>
<th>New Strategy (Switching To)</th>
<th>Old Strategy (Switching From)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Familiar or Rote</td>
</tr>
<tr>
<td></td>
<td>Unfamiliar or Deliberative</td>
</tr>
<tr>
<td>Familiar or Simple</td>
<td>High breaking cost.</td>
</tr>
<tr>
<td></td>
<td>Low acquiring cost.</td>
</tr>
<tr>
<td>Unfamiliar or Complex</td>
<td>Low breaking cost.</td>
</tr>
<tr>
<td></td>
<td>High acquiring cost.</td>
</tr>
</tbody>
</table>

Table B.2: The familiarity and micro-complexity of the old and new strategy interact to determine the cost of switching strategies.

For simplicity and concision, this discussion will not break down monitoring or switching costs further. If such additional specificity proves useful, CHSE can be easily expanded in future.

Each management cost component (monitoring and switching costs) for a given strategy will be the sum of such costs for each transition from that strategy to another strategy (each link away from the node in question to another node in figure B.9).

And the need to monitor will be attenuated by the frequency or probability of needing to switch to another strategy. If one never needs to switch, then one never needs to worry about detecting transition criteria.

Total monitoring cost is simply the sum of these costs, for all possible transitions from this strategy. Let \( \omega(s_j) \) be total monitoring cost for strategy \( s_j \), and \( \omega(t(s_j, s_i)) \) be the monitoring cost of a specific transition (a link in the network) from strategy \( s_j \) to \( s_i \), and \( P(t(s_j, s_i)) \) be the probability of such a transition occurring.

\[
\omega(s_j) = \sum_{i=1}^{n} \omega(t(s_j, s_i))P(t(s_j, s_i)) 
\] (B.10)

If the need to switch for individual transitions cannot be obtained, then the monitoring cost for the strategy could be estimated using the strategy’s overall need to switch to other strategies across the entire SEG (shown here as \( P(s_j') \), which is
the probability of the applied strategy in any randomly-selected eco-cell of the SEG being not $s_j$, but some other strategy):

$$\omega(s_j) = P(s'_j) \sum_{i=1}^{n} \omega(t(s_j, s_i))$$  \hspace{1cm} (B.11)

Switching cost would be calculated in much the same way. Let $\gamma(s_j)$ be total switching cost for strategy $s_j$, and $\gamma(t(s_j, s_i))$ be the switching cost of a specific transition (a link in the network) from strategy $s_j$ to $s_i$, and $P(t(s_j, s_i))$ be the probability of such a transition occurring.

$$\gamma(s_j) = \sum_{i=1}^{n} \gamma(t(s_j, s_i)) P(t(s_j, s_i))$$  \hspace{1cm} (B.12)

Or it could be estimated using the probability $P(s'_j)$, of not applying strategy $s_j$, as above.

$$\gamma(s_j) = P(s'_j) \sum_{i=1}^{n} \gamma(t(s_j, s_i))$$  \hspace{1cm} (B.13)

This requires a way of determining the probability of the strategy being applied (or not applied). This can be calculated as the strategy’s area of applicability – the percentage of the possibility space of the SEG that is covered in the corresponding colour of the strategy in question.

The broadness (and commonness) of the domain of application is a significant factor determining the actual ultimate utility of a reasoning strategy. However, area of applicability is a secondary factor in a strategy’s utility, because it is moderated by the cost of switching strategies. For example, if the cost of switching was zero, then there would be no reason not to switch, and the area of applicability would not be a factor at all in strategy utility. Similarly, if the area of applicability was 100% of the system, then that would also eliminate the cost of strategy management. To handle such variability, one must be able to separate out the broadness of the area of applicability component of the strategy’s utility (figure B.10).
One simple way to capture area of applicability is to look at the SEG, and tally up all the appearances of the same strategy. This entails finding the set of situations in the possibility space where the strategy in question is applied. Define a new set $\zeta(s)$ containing all situations $\varsigma$ in the entire set $\zeta$ where the strategy applied in that situation $s(\varsigma)$ is the strategy in question $s$.

\[
\zeta(s) = \varsigma \in \zeta | s(\varsigma) = s
\]  

(B.14)

The area of applicability (or the probability of the strategy occurring) is not simply the sum of these situations. Each situation should be weighted by its probability. For example, a strategy may be dominant in a large number of individual situations, but all those situations might be negligibly rare. Another strategy might only appear in a small number of situations, but those situations are extremely common. The variation in frequency or probability of different situations in a system can be represented visually by the saturation or brightness of pixels (figure B.11), which can be overlaid on the SEG (figure B.8).
To calculate this area of applicability, one must first gather together the situations where the strategy is viable, and then go through the list and sum the probabilities of these situations (where the probability of a situation describes the percentage of the time that strategy is likely to occur when dealing with the system). The resulting total describes what percentage of the time engaging with the system that a user can spend using this strategy. Let $P(s)$ be the broadness of the area of applicability of strategy $s$ (its probability of being applied), $\varsigma$ be a situation in the set $\zeta(s)$ defined above (i.e. situations where the strategy is applied), and $P(\varsigma_i)$ be the probability of such a situation occurring.

$$P(s) = \sum_{i=1}^{n} P(\varsigma_i) \tag{B.15}$$

Note that it must be limited to the situations where the strategy is viable, because summing all of probabilities of all situations would always total to 100%. As noted previously, one could apply an infinite number of strategies to all situations. But the vast majority would perform about as well as guessing, and thus would not be worth their computation cost. Therefore, one needs some criteria by which one decides a strategy is not viable for a situation. One of the most basic requirements is that the strategy performs better than chance. Another is to ensure the strategy’s baseline utility is positive – that its benefits outweigh its computation costs. These are the most basic of inclusion criteria for viability, but future work might justify more sophisticated and discriminating criteria.
A more fine-grained measure could be achieved by finding a continuous measure of a strategy’s viability in a situation, and weighting each situation by the specific degree of viability of the strategy in that situation. Basic utility, as calculated above in the *Calculating Utility* section, could serve such a role. The frequency of each situation could then be weighted by the strategy’s usefulness (i.e. its utility as calculated) in that situation, to ensure that situations where the strategy is more prominent are given more weight in calculating its area of applicability. Note that this process, by including utility, also gives a measure of simply the overall utility (or dominance) of a strategy. Let $\beta(s)$ be the generally predicted basic utility of a strategy, $P(\varsigma_i)$ be the probability of a situation where the strategy is viable, and $\beta(s, \varsigma_i)$ be the basic utility of the strategy in that situation.

$$\beta(s) = \sum_{i=1}^{n} P(\varsigma_i) \beta(s, \varsigma_i) \quad (B.16)$$

In order for this to be valid, the cost of switching will have to be measured on the same scale as the baseline strategy utility (including the benefits and the costs). For example, if strategy benefit in a study is measured in how many dollars the participant will receive at the end of the experiment, then the computation cost and the switching cost both need to be converted to dollar values to be put on the same scale, before they can be added or subtracted from each other. This will undoubtedly be an easier task for studies in computer science and artificial intelligence, where cost simply equals computation required, and is therefore a precise, knowable quantity.

These calculations could be used to take any strategy that appears on the viable strategy graph, or the entire possibility space, and adjust its baseline utility by its overall versatility value. Doing so will incorporate the area of applicability into the utility of the strategy and therefore determine if, for example, the broadness of its applicability is enough to make it a dominant strategy. It will also determine if the narrowness of a strategy’s area of applicability overpowers its baseline utility to render it an inefficient strategy that will not be used. This will solve the problem of the simplifying assumptions made in my original description of how to construct
a viable strategy graph. Those simplifying assumptions did not take into account the value of broadness of area of applicability of a strategy, and so may have pruned away strategies that might turn out to dominate. By performing this adjustment to strategy utility to incorporate versatility value, such dominant strategies should rise to the top of the utility column in a very clear and obvious manner. Therefore, this adjustment also provides precise testable predictions on if, when, and which strategies will come to dominate due to area of applicability.

B.5 CHSE Summary

It is difficult to interpret the somewhat mixed results in cognitive science about learning complex systems (Kahneman and Klein, 2009; Meadows and Wright, 2008; Richardson and Norgate, 2014; Rouwette, Größler, and Vennix, 2004; Sterman, 2000). Yet such cognitive studies investigating complex systems rarely employ any measures of complexity from complexity theory. This could be because most complexity theory work is too abstract to have obvious application to cognition. An overarching framework could specify how exactly various theories of complexity and cognition can be brought together to explain the experience of emergence in complex systems. CHSE conceptualises an emergent system as a cognitive habitat that gives rise to a diverse ecosystem of reasoning strategies. Within CHSE, emergence is the condition where one reasoning strategy outperforms another in terms of costs vs. benefits, derived from computation vs. accuracy. The model’s assumptions, functionality, and possible ways to measure relevant variables were discussed, along with a range of interesting implications and testable predictions for cognition.

B.5.1 Predictions

CHSE claims that sometimes, a learner’s mental model does not increase in comprehensiveness and cohesion, but actually splits up into multiple specialised modules that can grow more distinct and divergent as they develop. A new strategy will be adopted when its overall utility is greater than the previous strategy, conceptualised
in terms of the prediction bang for the computation buck. Once we have developed ways to measure the variables involved, we would then be able to calculate exactly when this would occur using the procedures described above. Once it has occurred, one would expect largely distinct reasoning to occur, meaning that the old and the new strategy should have relatively independent learning gains. Such gains could take the form of improved accuracy and / or reduced computation costs, for example through automatisation.

Various experimental manipulations should be able to promote such a strategy transition (i.e. emergence), by acting through variables that are either endogenous or exogenous to the system that participants are learning. Adding variables and interactions to the system would reduce the utility of a full simulation of a completely isomorphic mental model, thereby improving the viability of simpler reasoning strategies (e.g. heuristics). Similarly, making the system sensitive to initial conditions increases the chances of errors accumulating when attempting to predict several steps ahead in time, again reducing the effectiveness of a complete mental simulation. An example of this effect could manifest as a separation of implicit and explicit learning, if the explicit learning is (as is often the case) an isomorphic mental model of a system that is impractical to fully mentally simulate. This predicts that emergent systems are likely to cause implicit and explicit learning to occur independently of each other, whereas simple systems and likely to tightly couple implicit and explicit learning.

Introducing time restrictions (either endogenously or exogenously) will reduce the quantity of acceptable computation by increasing the need for quick thinking. Lowering the acceptable upper limit on computation is likely to affect which reasoning strategies are viable. Other ways to reduce available cognitive resources include occupying participants’ minds with additional parallel tasks or distractions.

Prediction bang could be manipulated by providing participants with better information or by correcting a misconception, thereby increasing the accuracy of predictions with a strategy. Alternatively, the computation cost of a strategy could be reduced by giving participants practice to better automatise the strategy. Either
approach will increase the overall utility of a strategy, potentially to the point where it outperforms a previously dominant strategy.

A more direct approach would be to manipulate the actual benefits and costs of strategies. For example, if only accuracies 95% and higher result in any reward, then that will promote very different strategies than the same system but with equal rewards given for any accuracy better than 55%. Alternatively, one could keep the accuracy threshold the same, and change the magnitude of the rewards and penalties.

A system with a more fully populated viable strategy graph will be more prone to developing a singular or small number of broadly-applicable dominant strategies, and therefore will have lesser macro-complexity, and therefore will be experienced as less complex than a system with a sparser viable strategy graph.

Manipulations that cause a system to have a slightly less open and populated viable strategy graph will not necessarily result in a significant increase in macro-complexity, but may simply reduce the utility of the small number of versatile dominant strategies. It depends if the bump downward in utility is equal to or greater than the cognitive cost of strategy switching.

A system with greater macro-complexity will support a greater diversity of reasoning strategies and will consequently be experienced as more complex than a system with less macro-complexity.

A system with greater temporo-macro-complexity will require more frequent adaptation and refinement of strategies, including refining the boundaries of strategies’ areas of applicability, and potentially the need to create new strategies to deal with situations that arise.

Available reasoning strategies can be considered as nodes in a network, connected by links denoting the conditions that justify switching strategies. CHSE implies that a person is likely to occupy a node on the network with some degree of inertia. Rather than constantly check the relevance of all possible strategies, one is likely to continue with a strategy and wait until a salient cue signals that a switch is
necessary or advantageous. With past experience of the system, such a cue might be a recognised set of conditions that demarcate the boundary of the current strategy’s area of applicability. But if it is a novel system, then that salient cue will likely take the form of a surprising event – a failure of the strategy to yield accurate predictions. Waiting for such a cue is often more efficient than constantly re-evaluating the relevance of all possible strategies.

Maintaining a large list of possible transitions from any one strategy is likely to be cognitively taxing. To keep more cognitive resources available for actually computing the strategy, one is likely to try to minimise the number of possible transitions from a strategy (i.e. minimise the number of links coming from a node in the strategy network). Consequently, CHSE predicts that a strategy network will tend to string strategies in series to create long chains with few branching points, where possible.

However, this effect is likely modulated by the micro-complexity of individual strategies. A strategy (node) that is relatively simple leaves available more cognitive resources that can be devoted to monitoring for cues that signal the need for a strategy transition – The node is able to support more links to other nodes. More complex strategies will not leave many cognitive resources for transition monitoring. However, even if ample cognitive resources are available, one is still likely to try to minimise cognitive effort, and thus still try to minimise the links coming from any node.

Finally, CHSE predicts that the local, raw utility of a strategy in a situation can be overpowered by versatility value. For example, a strategy may be optimal for one situation, but if the utility gained by switching is lesser than the cost of switching to this strategy, then switching is not the optimal choice. Thus, strategies that have a broad area of applicability have an inflated overall utility, making them more likely to dominate. I proposed methods to calculate this kind of versatility value if and when reliable measures for the relevant variables are developed.
B.5.2 Future Work on CHSE

Some of the predictions and claims described have promising corresponding findings in existing literature, but most will need to be empirically tested in future studies. Additionally, there are many areas in which CHSE should be elaborated and expanded. Possible ways to measure or quantify relevant variables were suggested throughout the explanation of CHSE, but none in particular are obviously ideal at this stage. Therefore, future work should propose and investigate the validity of different ways of measuring and quantifying the variables in the model of CHSE.

CHSE would benefit from studies developing objective measures for strategy difference or degree of emergence and comparing these measures with participants’ subjective reports of degree of emergence. This work will be necessary to determine which objective measures best capture what makes a strategy qualitatively different, and also to determine if there is a threshold of degree of emergence that people consistently report as being the threshold at which a phenomenon becomes significantly emergent. Utility difference, output difference, and different forms of content difference may prove useful here, or entirely different measures may need to be developed.

A related open question is how to quantify the computation cost of strategies for humans. Future studies will need to investigate the relationship between indicators of cognitive effort (e.g. time taken to decide, Likert scale ratings, or price willing to pay to avoid cognitive effort) and the nature of the strategies themselves, to find if certain measures of micro-complexity from information theory correspond to humans’ experience of strategy complexity.

Criteria were suggested to decide if a strategy is viable in a situation, and thus worthy of inclusion in a viable strategy graph, a strategy ecosystem graph, or versatility value calculations: the strategy performs better than chance, and / or the strategy’s local raw utility is positive. Future work could look into explaining and justifying more sophisticated criteria, or even proposing a continuous measure of viability. For the latter, I suggested that local raw utility could serve such a function, but a dedicated viability measure is certainly worth considering. Similarly, there
would be value in developing precise definitions of a ‘situation’: consistent, standardised procedures for categorising the possibility space into meaningfully-related sections to treat as a singular situation for the purposes of analysing strategy dominance.

CHSE could also be improved by a better understanding of the costs of strategy switching. Some factors were suggested, such as the micro-complexity of the switching criteria/conditions, the perceptual subtlety of switching cues, the number of possible transitions, and the actual act of switching by breaking the current mental routine. Future work may suggest additional factors, or ways to quantify them. Attention was mentioned as a factor, but attention could be more fully incorporated into CHSE on a more fundamental level, perhaps in relation to computation costs.

While CHSE provides certain predictions on how and when adaptation and strategy switching will occur, there are still many unresolved questions. Temporomacro-complexity (the dynamics of a changing strategy ecosystem) should be further elaborated in order to generate specific predictions for strategy switching and adaptation. Even without the temporal component, there are still many unanswered questions: When exactly is it appropriate to use a surprising event to merely adjust one’s understanding of a strategy’s accuracy, or to adjust the boundaries of applicability for that strategy, or to try to expand and refine that strategy to explain both the old and the new data? And how exactly does one go about constructing a new strategy or modifying an existing one?

Finally, work is ongoing to formally integrate CHSE with the MMT framework where it can provide more detail on the process of mastery.

B.6 Conclusion

Cognitive studies have investigated people’s ability to handle complex systems and situations, but no standard conception of complexity has been adopted from the complexity theory literature. This could be due to the fact that most of that literature
is not focussed on generating testable predictions about human cognition, but more on very technical and abstract mathematical discussions. Here, I sought to address this by proposing a model that can function as an integrating framework that specifies how different theories of complexity and cognition can be brought together to explain how a system is experienced as emergent. The model of emergence I proposed, CHSE, was specifically built around its implications for human cognition. But CHSE also has relevance to related domains such as artificial intelligence and machine learning.

CHSE models an emergent system as a cognitive habitat capable of supporting a diverse ecosystem of competing reasoning strategies. I detailed the structure of this model and consequent predictions for cognition and learning within a complex system. I used CHSE to provide new interpretations of some previous studies of human cognition, and in turn used such studies to inform the CHSE model. CHSE not only provides specific predictions, but also describes relations and interactions between disparate concepts in cognitive science, such as causal learning, cognitive load theory, automatisation, attention, dual process theory, and adaptation to surprising events.

CHSE can provide considerably more detail on the process of mastery than what was described in previous chapters detailing the MMT framework. This formal description of cognitive habitats and strategy ecosystems lays the groundwork for formally integrating CHSE into MMT to augment the framework.

There are many promising directions for future work in relation to CHSE, not just in terms of testing the many predictions it generates, but in expanding and elaborating the model to provide better tools with which to study cognition. By building a model of complexity specifically to generate testable predictions of human cognition, and by explaining the model in more concrete terms than the average complexity theory paper, CHSE should prove much easier to incorporate into cognitive research than many preceding definitions of complexity.
Appendix C

Ethics Clearance

All research was conducted within the ethics guidelines of Griffith University. All research that fell within the scope of requiring ethics review was approved.

GU Ref No: 2017/922

GU Ref No: 2019/387
Appendix D

PANAS Mood Scale Measurement Tool

PANAS is a mood measurement tool. Participants are presented with the following text:

This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you felt this way while playing this game. Use the following scale for your answers.

1 = very slightly or not at all
2 = a little
3 = moderately
4 = quite a bit
5 = extremely

Participants are then presented with the following items to rate on that scale of one to five (arranged in a random order for each participant to counter-balance for any ordering effects):

1. Interested
2. Distressed
3. Excited
4. Upset
5. Strong
6. Guilty
7. Scared
8. Hostile
9. Enthusiastic
10. Proud
11. Irritable
12. Alert
13. Ashamed
14. Inspired
15. Nervous
16. Determined
17. Attentive
18. Jittery
19. Active
20. Afraid
Appendix E

PENS Game Enjoyment Measurement Tool

PENS is a Likert-style measurement tool for assessing the quality of a game according to player ratings of several attributes of the game experience, on a scale of one to seven.

PENS is a copyrighted tool employed in the games industry and therefore cannot be reproduced here without permission. Those looking to obtain a copy should contact the owners of the copyright via www.immersyve.com
Appendix F

Coding Instructions For Experiment 2.1

What participants wrote when asked about the rules of the flocking algorithm needed to be converted into a numerical value for statistical analysis. Below are the instructions that the two blinded coders were given for converting the dynamic complexity explicit knowledge text responses of participants in experiment 2.1:

Participants were asked the question:

"What are the rules determining how the creatures move?"

There are three rules and two points per rule, so a participant should score a maximum of six points when they accurately describe all three rules, and zero points when they accurately describe none.

The creatures moved using a standard flocking algorithm. The rules were:

1. Cohesion: Steer towards average position of neighbours (long range attraction).

Participants had three text boxes they could use to type answers. A fully correct set of responses might be: 1. Move towards the centre of mass of the others.
2. Try to match the average velocity of the others to all head in a similar direction. If you get too close to another individual, move away to avoid collision.

But they may be worded differently, contain fewer details, or several different rules may be described in a single line of text. Or, one rule might be split across multiple lines. For example, a participant may have made two observations that they wrote in two different lines, and individually they are simple and incomplete, but when taken together describe one or more of the rules above. It is not necessary for the three rules to each be in separate lines, and it doesn’t matter in which order they appear (i.e. cohesion doesn’t need to be the first rule described).

Note that the text fields where participants entered their answers didn’t support any special characters or punctuation, replacing them with a lower-case letter “x”, so that “Hello!” becomes “Hellox” and “don’t” becomes “donxt”. Ignore this and any other spelling or grammatical errors.

Rule Completeness: 1 point per rule
A full point should be awarded if a rule is accurately described. A half point can be awarded if the response correctly identifies a relationship or set of relevant variables, but is vague or ambiguous, and thus incomplete. If they do not describe the rule at all, they receive zero points.

Rule Attribution: 1 point per rule
The player’s avatar was slightly larger than the others in the flock, and it was controlled by the mouse. So if they refer to the larger creature, or the mouse movement, they are assumed to be referring to the player’s avatar. The player’s avatar was treated as a member of the flock, and thus the rest of the flock reacted to the player as they would their neighbours. Therefore, if a participant described one of the above rules as applying to the player, without also stating that it applies equally to the others in the flock as well, then they are half-right and receive half a point. If they describe the rule as applying to the whole flock, or the creatures generally, then they receive a full point. If they didn’t get at least half a point for completeness, then they can’t get any points for attribution.
The player’s avatar was slightly larger than the others in the flock, and it was controlled by the mouse.

Example of how a set of responses might be coded:

**Example Coding Scores for a Participant**

<table>
<thead>
<tr>
<th>Rule</th>
<th>Completeness. Correctly describes the rule = 1. Incomplete rule (correctly identifies relationships or variables but is vague or ambiguous) = 0.5</th>
<th>Attribution. Attributes the rule to the whole flock = 1. Attributes the rule only to the player’s avatar = 0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohesion</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Alignment</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Separation</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The coder’s job is to assign points for the completeness and attribution of the three rules.

Cohesion, Completeness: 1 full point if they say that the creatures are attracted to each other. Half a point if they correctly identify variables or relationships (proximity attraction), but they are vague or ambiguous.

Cohesion, Attribution: 1 full point if they say that the creatures are attracted to each other. Half a point if they say that the creatures were attracted only to the player, without also stating that they were attracted to each other.

Alignment, Completeness: 1 full point if they say that the creatures try to move in the same direction. Half a point, as above, if they correctly identify variables or relationships (eg direction, velocity), but they are vague or ambiguous.

Alignment, Attribution: 1 full point if they say that the creatures try to match each other’s heading. Half a point if they say that the creatures try to match only the player’s heading.

Separation, Completeness: 1 full point if they say that the creatures are repelled or try to avoid collision or maintain a certain distance from each other. Half a point, as above, if they correctly identify variables or relationships (eg proximity repulsion), but they are vague or ambiguous.
CODING INSTRUCTIONS FOR EXPERIMENT 2.1

Separation, Attribution: 1 full point if they say that the creatures are repelled from each other. Half a point if they say they are repelled from only the player.

Examples:

Participant 1 wrote:

“They are affected by each others’ velocities”

Participant 1 score: 0.5 points for Alignment-Completeness because they didn’t specify how they were affected. 1 point for Alignment-Attribution because they correctly stated that it is a mutual effect between all flock members, not just the player.

If a response is irrelevant (e.g. “I like pie”), then treat it as blank.
Appendix G

Scale-Free Network Algorithm

Pseudocode

The scale-free networks in study two were generated via code with the following structure:

```csharp
public int SelectNode()
{
    int tempNodeNum = Random(0, nodes.Length);
    int tempLinkNum = Random(0, links.Length+1);
    if(tempLinkNum < links.Length)
    {
        tempNodeNum = links[tempLinkNum].GetRandomConnectedNode();
    }
    return tempNodeNum;
}

public void GenerateScaleFreeNetwork()
{
    int tempFrom = Random(0, nodes.Length);
    int tempTo = Random(1, nodes.Length); //Start from 1, because nothing
    AddLink(tempFrom, tempTo);
}
```
for(int i = 0; i < nodes.Length; i++)
{
    tempFrom = i;
    tempTo = SelectNode();
    if(tempTo == 0) //If it selected the player, re-select another,
        because nothing can activate the player
        {
            tempTo = Random(1, nodes.Length);
        }
    AddLink(tempFrom, tempTo);
}
for(int i = 1; i < nodes.Length; i++)
{
    tempFrom = SelectNode();
    tempTo = i;
    AddLink(tempFrom, tempTo);
}
for(int i = 0; i < 8; i++)
{
    tempFrom = SelectNode;
    tempTo = SelectNode();
    if(tempTo == 0) //If it selected the player, re-select another,
        because nothing can activate the player
        {
            tempTo = Random(1, nodes.Length);
        }
    AddLink(tempFrom, tempTo);
}

int tempFrom = 0; //Add a link from the player to ensure the player
        can start off the chain of activation.
    int tempTo = Random(1, nodes.Length);
    AddLink(tempFrom, tempTo);
Appendix H

Lyapunov Calculation Pseudocode

The Lyapunov exponent in study two was calculated with code with the following structure:

```java
private double initialLyaDist = 1.0; //What distance constitutes a "nearby" condition (i.e. initialLyaDist) is a somewhat arbitrary judgement call based on observation of the system. This study used an initialLyaDist of 1.
private int stateVecLength = Math.RoundToInt(boidDensity * numVarsPerBoid);
private int testStateVecLength = Math.RoundToInt(boidDensity * 2.0);
    //Two values per boid: One value for the boid’s distance to average centre of the flock, and one value for the boid’s distance to average velocity of the flock.
private static int numVarsPerBoid = 4;
private int lyaMeasureTime = 100; //The system must be simulated forward in time a certain amount in order to see if the nearby starting conditions diverge or converge. The decision of exactly how far to simulate forward before getting a measurement (i.e. the value of lyaMeasureTime) is a judgement call based on observation of the system. This study used a lyaMeasureTime of 100.
private int lyaTimeStep = 40; //How long the system would wait before measuring again (by generating variations and simulating them forward to see how much they diverged)
currentTestVec = new VectorN(testStateVecLength);
```
LYAPUNOV CALCULATION PSEUDOCODE

9 nearbyTestVec = new VectorN[stateVecLength]; //Number of nearby
      conditions simulated = the number of dimensions to describe the
      system
10 initialTestDist = new double[stateVecLength];
11 startingNearbyVec = new VectorN[stateVecLength];
12 pastNearbyVec = new VectorN[stateVecLength];
13 currentNearbyVec = new VectorN[stateVecLength];
14 for(int i = 0; i < startingNearbyVec.Length; i++)
15 {
16   startingNearbyVec[i] = new VectorN(stateVecLength);
17   currentNearbyVec[i] = new VectorN(stateVecLength);
18   pastNearbyVec[i] = new VectorN(stateVecLength);
19   nearbyTestVec[i] = new VectorN(testStateVecLength);
20   initialTestDist[i] = 0.0;
21 }

22 public void StartCalculatingLya()
23 {
24   calculatingLya = true;
25   lyaMeasureTimer = lyaMeasureTime;
26   currentTrajectory = -1;
27   InitialiseStateVec();
28 }

29 public void StopCalculatingLya()
30 {
31   calculatingLya = false;
32   lyaMeasureTimer = lyaMeasureTime;
33 }

34 public void InitialiseStateVec()
35 {
36   UpdateStateVec();
37   InitialiseTrajectories();
38 }

39 public void InitialiseTrajectories() //Define several nearby initial
      conditions as being a certain distance, "initialLyaDist", from zero
      in several dimensions. A sample of possible nearby conditions is

351
used because there is insufficient computation to check every
single possible nearby condition.

```csharp
for (int i = 0; i < startingNearbyVec.Length; i++)
{
    for (int k = 0; k < stateVecLength; k++)
    {
        if (k == i)
        {
            startingNearbyVec[i].SetElement(i, initialLyaDist);
        }
        else
        {
            startingNearbyVec[i].SetElement(k, 0.0);
        }
    }
}

for (int i = 0; i < startingNearbyVec.Length; i++)
{
    currentNearbyVec[i].SetVector(startingNearbyVec[i]);
}

public VectorN TakeVecSnapshot()
{
    for (int i = 0; i < boids.Count; i++)
    {
        int k = i * numVarsPerBoid;
        Boid tempBoid = (Boid)boids[i];
        if (tempBoid != null)
        {
            Vector3d tempPos = tempBoid.GetCurrentPos();
            Vector2d tempVel = tempBoid.GetCurrentVel();
            snapshotVec.SetElement(k, tempPos.x);
            snapshotVec.SetElement(k + 1, tempPos.z);
            snapshotVec.SetElement(k + 2, tempVel.x);
            snapshotVec.SetElement(k + 3, tempVel.y);
        }
    }
```
public VectorN CheckTestStateVec()
{
    Vector2d tempAvgPos = new Vector2d(0.0, 0.0);
    Vector2d tempAvgVel = new Vector2d(0.0, 0.0);
    for (int i = 0; i < boids.Count; i++)
    {
        DynAgent tempAgent = (DynAgent)boids[i];
        tempAvgPos += tempAgent.GetCurrentPos2d();
        tempAvgVel += tempAgent.GetCurrentVel();
    }
    tempAvgPos /= boids.Count;
    tempAvgVel /= boids.Count;
    for (int i = 0; i < boids.Count; i++)
    {
        Boid tempBoid = (Boid)boids[i];
        Vector2d tempPos = tempBoid.GetCurrentPos2d();
        Vector2d tempVel = tempBoid.GetCurrentVel();
        testPosVec.SetElement(i, Vector2d.Distance(tempPos, tempAvgPos));
        testVelVec.SetElement(i, Vector2d.Distance(tempVel, tempAvgVel));
    }
    testPosVec.Sort();
    testVelVec.Sort();
    for (int i = 0; i < boids.Count; i++)
    {
        int k = i * 2;
        testStateVec.SetElement(k, testPosVec.GetElement(i));
        testStateVec.SetElement(k + 1, testVelVec.GetElement(i));
    }
    return testStateVec;
}
```java
public void UpdateStateVec()
{
    if(currentTrajectory < 0)
    {
        startingStateVec.SetVector(currentStateVec);
        currentStateVec.SetVector(TakeVecSnapshot());
        currentTestVec.SetVector(CheckTestStateVec());
        ApplyNearbyConditions();
    }
    else
    {
        currentNearbyVec[currentTrajectory].SetVector(TakeVecSnapshot());
        nearbyTestVec[currentTrajectory].SetVector(CheckTestStateVec());
    }
}

public void AdvanceTrajectory()
{
    UpdateStateVec();
    currentTrajectory++;
    if(currentTrajectory >= stateVecLength)
    {
        currentTrajectory = -1;
    }
}

public void ApplyNearbyConditions()
{
    for(int i = 0; i < startingNearbyVec.Length; i++)
    {
        currentNearbyVec[i].SetVector(startingNearbyVec[i]);
        currentNearbyVec[i].Add(currentStateVec);
    }
}

public void MoveTrajectoriesToOrigin()
{
for (int i = 0; i < startingNearbyVec.Length; i++)
{
    currentNearbyVec[i].Subtract(currentStateVec);
}

public void ConvertTrajectoriesToNearbyConditions()
{
    currentNearbyVec = GramSchmidt(currentNearbyVec);
    for (int i = 0; i < startingNearbyVec.Length; i++)
    {
        currentNearbyVec[i].Normalise();
        currentNearbyVec[i].Scale(initialLyaDist);
        startingNearbyVec[i].SetVector(currentNearbyVec[i]);
    }
}

public VectorN GetTestStateVec(VectorN _vec)
{
    VectorN tempAvgPosVec = new VectorN((int)boidDensity);
    VectorN tempAvgVelVec = new VectorN((int)boidDensity);
    Vector2d tempAvgPos = new Vector2d(0.0, 0.0);
    Vector2d tempAvgVel = new Vector2d(0.0, 0.0);

    for (int i = 0; i < boidDensity; i++)
    {
        int k = i*GetNumVarsPerBoid();
        Vector2d tempPos = new Vector2d(0.0, 0.0);
        Vector2d tempVel = new Vector2d(0.0, 0.0);
        tempPos.x = _vec.GetElement(k);
        tempPos.y = _vec.GetElement(k+1);
        tempVel.x = _vec.GetElement(k+2);
        tempVel.y = _vec.GetElement(k+3);
        tempAvgPos += tempPos;
        tempAvgVel += tempVel;
    }
    tempAvgPos /= boidDensity;
tempAvgVel /= boidDensity;

for(int i = 0; i < boidDensity; i++)
{
    int k = i*GetNumVarsPerBoid();
    Vector3d tempPos = new Vector2d(0.0, 0.0);
    Vector2d tempVel = new Vector2d(0.0, 0.0);
    tempPos.x = _vec.GetElement(k);
    tempPos.z = _vec.GetElement(k+1);
    tempVel.x = _vec.GetElement(k+2);
    tempVel.y = _vec.GetElement(k+3);
    tempAvgPosVec.SetElement(i, Vector2d.Distance(tempPos, tempAvgPos));
    tempAvgVelVec.SetElement(i, Vector2d.Distance(tempVel, tempAvgVel));
}

tempAvgPosVec.Sort();
tempAvgVelVec.Sort();

VectorN tempStateVec = new VectorN(testStateVecLength);
for(int i = 0; i < boidDensity; i++)
{
    int k = i * 2;
    tempStateVec.SetElement(k, tempAvgPosVec.GetElement(i));
    tempStateVec.SetElement(k+1, tempAvgVelVec.GetElement(i));
}
return tempStateVec;

public void SetWorldState()
{
    if(currentTrajectory < 0)
    {
        SetWorldState(currentStateVec);
    }
    else
    {
        SetWorldState(currentNearbyVec[currentTrajectory]);
```java
public void CalculateLya()
{
    AdvanceTrajectory();

    if(lyaMeasureTimer < lyaMeasureTime && currentTrajectory < 0)
    {
        if(startingStateVec != null)
        {
            MoveTrajectoriesToOrigin();

            int longestVec = -1;
            double largestAdjMag = 0.0;

            for(int i = 0; i < nearbyTestVec.Length; i++)
            {
                nearbyTestVec[i].Subtract(currentTestVec); //Moving to the origin
                double tempAdjMag = GetTestVecMagnitude(startingNearbyVec[i]);
                tempAdjMag = nearbyTestVec[i].GetMagnitude() / tempAdjMag;
                if(tempAdjMag > largestAdjMag)
                {
                    longestVec = i;
                    largestAdjMag = tempAdjMag;
                }
            }

            lyaNum = largestAdjMag;
            WriteString(lyaDataPath, lyaNum.ToString() + " "); //Saving a recording of the current Lyapunov number to a file
            ConvertTrajectoriesToNearbyConditions();
        }
    }

    SetWorldState();
}
```
if (currentTrajectory < 0) {
    lyaMeasureTimer--; 
    if (lyaMeasureTimer <= 0) {
        StopCalculatingLya();
    }
}
lyaStepTimer = lyaTimeStep;


Boyle, Elizabeth, Thomas M. Connolly, and Thomas Hainey (2011). “The role of psychology in understanding the impact of computer games”. In: *Entertainment


Canossa, Alessandro, Josep B. Martinez, and Julian Togelius (2013). “Give me a reason to dig Minecraft and psychology of motivation”. In: *IEEE Conference on Computational Intelligence and Games, CIG*. ISSN: 23254270. DOI: 10.1109/CIG.2013.6633612.


Cook, Claire, Noah D. Goodman, and Laura E. Schulz (2011). “Where science starts: Spontaneous experiments in preschoolers’ exploratory play”. In: Cognition 120.3,


Cummings, Jj and Tl Ross (2011). “Optimizing the psychological benefits of choice: Information transparency and heuristic use in game environments”. In: *Internat-


DOI: 10.1080/0950069980201002.


Dunbar, Norah E. et al. (2014). “Implicit and explicit training in the mitigation of cognitive bias through the use of a serious game”. In: Computers in Human Behavior 37, pp. 307–318. ISSN: 07475632. DOI: 10.1016/j.chb.2014.04.053. URL: http://dx.doi.org/10.1016/j.chb.2014.04.053.


DOI: 10.1177/073194871003300103. URL: http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=f0a33cc1-f16c-49ac-bd2b-e8f3149fa699@sessionmgr4004&\&vid=9\&hid=4212.


Eseryel, Deniz et al. (2012). “Dynamic Modeling as a Cognitive Regulation Scaffold for Developing Complex Problem-Solving Skills in an Educational Massively
Multiplayer Online Game Environment”. In: *Journal of Educational Computing Research* 45.3, pp. 265–286. ISSN: 0735-6331. DOI: 10.2190/EC.45.3.a.


Fraser, James, Michael Katchabaw, and Robert E. Mercer (2014). “A methodological approach to identifying and quantifying video game difficulty factors”. In: *Entertainment Computing* 5.4, pp. 441–449. ISSN: 18759521. DOI: 10.1016/j.entcom.2014.08.004. URL: http://dx.doi.org/10.1016/j.entcom.2014.08.004.


— (2008). “Game-like learning: An example of situated learning and implications for opportunity to learn”. In: Assessment, Equity, and Opportunity to Learn, pp. 200–221. DOI: http://dx.doi.org/10.1017/CBO9780511802157.010.


Green, C. Shawn, Alexandre Pouget, and Daphne Bavelier (2010). “Improved probabilistic inference as a general learning mechanism with action video games”. In:


388


Johnson, Soren (2013). When choice is bad- finding the sweet spot for player agency. URL: http://gamasutra.com/view/news/197181/When\_choice\_is\_bad\_finding\_the\_sweet\_spot\_for\_player\_agency.php (visited on 10/12/2013).


Kim, Min Kyu (2015). “Models of learning progress in solving complex problems: Expertise development in teaching and learning”. In: Contemporary Educational Psychology 42, pp. 1–16. ISSN: 10902384. DOI: 10.1016/j.cedpsych.2015.03.005. URL: http://dx.doi.org/10.1016/j.cedpsych.2015.03.005.


Konert, Johannes et al. (2014). “Modeling the Player: Predictability of the models of Bartle and Kolb based on NEO-FFI (Big5) and the implications for game based learning”. In: International Journal of Game-Based Learning 4.2, pp. 36–50. ISSN: 2155-6849. DOI: 10.4018/ijgbl.2014040103.


Kottonau, Johannes (2011). “An interactive computer model for improved student understanding of random particle motion and osmosis”. In: Journal of Chemical Education 88.6, pp. 772–775. ISSN: 00219584. DOI: 10.1021/ed100088h.


Liu, Peng and Zhizhong Li (2012). “Task complexity: A review and conceptualization framework”. In: *International Journal of Industrial Ergonomics* 42.6, pp. 553–


Mohsen, Mohammed Ali (2016). “The Use of Computer-Based Simulation to Aid Comprehension and Incidental Vocabulary Learning”. In: Journal of Educational


Oliver, Mary Beth et al. (2015). “Video Games as Meaningful Entertainment Experiences”. In: Psychology of Popular Media Culture.


Pearce, Celia (2002). Sims, battleBots, cellular automata God and go: A conversation with will wright.


Rigby, Scott and Richard Ryan Design (2007). “Sharpening Our Focus: Looking Deeper into the Measurement of the Player Experience”. In:


BIBLIOGRAPHY


Scheffler, I. (2009). “Science and subjectivity”. In: *Literary Paritantra (Syst)* 1.3, p. 4. DOI: EBNCSCHEF. URL: http://books.google.co.uk/books?hl=en\&lr={\&}id=cgqjFPCQfUsC\{\&}oi=fnd\{\&}pg=PR7\{\&}dq=scheffler+science\{\&}ots=XhcTUKxi1Vl\{\&}sig=Z5Cnhttp://books.google.co.uk/books?hl=en\&lr={\&}id=cgqjFPCQfUsC\{\&}oi=fnd\{\&}pg=PR7\{\&}dq=scheffler+science\{\&}ots=XhcTUKxi1Vl\{\&}sig=.


Sluis, Frans van der et al. (2014). “When Complexity Becomes Interesting”. In: Journal of the Association for Information Science and Technology 65.7, pp. 1478–1500.


Spielberger, C D et al. (1979). “Preliminary manual for the state-trait personality inventory (STPI)”. In: Unpublished manuscript University of South Florida Tampa.


Testa, Bernard (2009). “Dispersal (Entropy) and recognition (Information) as foundations of emergence and dissolvence”. In: Entropy 11.4, pp. 993–1000. ISSN: 10994300. DOI: 10.3390/e11040993.


Thompson, N., K. Peppler, and J. Danish (2010). “Designing BioSim: Playfully encouraging systems thinking in young children”. In:


Tovey, Mark (2013). Balance of the Planet — Simulation Game Evaluation. URL: http://socialinnovationsimulation.com/2013/06/18/simulation-game-analysis-balance-of-the-planet/ (visited on 08/19/2013).


ence museums”. In: *Science Education* 94.5, pp. 794–809. ISSN: 00368326. DOI: 10.1002/sce.20394.


Wilson, Timothy D. et al. (2014). “Just think: The challenges of the disengaged mind”. In: *Science* 345.6192, pp. 75–77. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1250830. URL: http://www.sciencemag.org/content/345/6192/75/5Cnhttp://www.sciencemag.org/content/345/6192/75.full/5Cnhttp://www.sciencemag.org/content/345/6192/75.full.pdf.


