Reflection in Creativity Support Tool Interaction: Characterisations for AI-based Music Composition

Corey Ford

Submitted in partial fulfilment of the requirements
for the Degree of Doctor of Philosophy

Centre for Digital Music

School of Electronic Engineering and Computer Science

Queen Mary University of London

Statement of Originality

I, Corey Ford, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below. I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I accept that Queen Mary University of London has the right to use plagiarism detection software to check the electronic version of the thesis. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature:

Submitted for Viva: 21st September 2024
Submitted with Corrections: 30th July 2025

Details of collaboration and publications: All research and contributions in this thesis and the associated publications are my own work. The research was supported by Prof Nick Bryan-Kinns within the scope of his role as my primary supervisor, and he is acknowledged as an author in all related publications. In Chapter 7, all work is my own excluding the first-person accounts and music compositions which were provided by co-authors. Previous publications related to this thesis are described in Section 1.6.

Abstract

Creativity Support Tool (CST) evaluations in Human-Computer Interaction predominantly apply mixed-methods to assess user engagement. For example, CST evaluations include questionnaires based on the theory of flow: an optimal state where people feel in control and lose self-awareness. Reflection is also crucial in CST interaction. However, reflection is marked by ambiguity and self-awareness, which contrasts with engagement. Few CST evaluations address reflection, and even fewer use reflection questionnaires in systematic mixed-methods evaluations. This thesis challenges the dominance of engagement-based evaluation in CSTs. It argues that a systematic evaluation of reflection in CST interaction is needed to characterise and support the user's creative process.

The thesis thus develops the Reflection in Creative Experience (RiCE) questionnaire to systematically evaluate reflection in CST interaction. RiCE is applied across user studies in the case study domain of composing music with Artificial Intelligence Generated Content (AIGC). The domain represents the state-of-the-art in CST development and provides a rich and specific context for characterising reflection. The focus is on generative AI: models built from datasets that produce new data with similar properties.

A user study with RiCE characterised reflection in artist-researchers' use of different AI tools; they reflected on their process when curating AIGC in real-time, and reflected on themselves when organising their curated AIGC. A new AI-based musical CST for reflection was also evaluated, characterising reflection's interplay with moments of focused engagement; during focus, reflection occurs when users are uninterrupted and learn from AIGC.

The new knowledge contributions are the RiCE questionnaire and novel characterisations of reflection in AI-based music composition. RiCE enables the systematic assessment of reflection in different CSTs and study conditions – an advance on existing engagement-focused tools. The new characterisations of reflection provide value to AI-based musical CST users by showing how to invoke different types of reflection in their practice.

Acknowledgements

Firstly, thanks are not just due but awarded lavishly in heaps to my supervisor Prof Nick Bryan-Kinns. Thank you for the excellent advice and guidance throughout the PhD, far beyond the job description! Thank you also to my progression panel. To Prof Simon Colton for encouraging critical thinking on AI, creativity and reflection. And to Dr Tony Stockman for the support not just in the PhD, but for the career advice and opportunities. Thank you also to my examiners Prof Ernest Edmonds and Prof Josh Reiss – it's an honour to be assessed by experts whose work I deeply admire.

Thank you to everyone in the AIM CDT who made my time at QMUL a friendly and enriching experience. In particular, to Alvaro Bort for the support above and beyond the call of duty. To my QMUL friends: Berker, Courtney, Sara, Ash, Jack, Pedro, Lizzie, Lewis, Ellie, Remi, Max, Harnick, Jingjing, Bleiz, Alex, Shuoyang, Sophie, Jack, David and others. To IGGI friends: Amy, Kyle, Terry and Sebastian. Supervision group friends: Ulfa, Jiali, Jianing, Teo, Nicole, Muhan and Antonella. My Creativity & Cognition friends: Jeba, Yinmiao, Rosa, Vlad and Amna. And friends outside QMUL: Tom, Laura, Sam, Iggy, Ollie & Hannah, Sophie, Ben and Callum.

To academics I've been lucky enough to work alongside: Katja Ivanova, Karen Shoop, Pat Healey, Tassos Tombros, Julian Hough, Alan Chamberlain, Alex Lerch, Charis Saitis and others. To my alma mater for the ongoing support: Chris Nash, Max Davis, Luke Child and others.

Thank you to my family for love and care: Mum, Dad, Freddie, Paige, Nanny Gill, Grandad John, Nanna Helen, Grandad Another John, Uncle Peter, Auntie Carol, Sam, Charlotte, Auntie Bernie, Uncle Dave and Luke. To my extended family: Teresa, John, Adam, Leslie, Barbara (Nanny Horse), Barbara (Nanny Brandy), Ray, Jim, Katie, and others.

Lastly, to Leilani Davis, for being at my side throughout the whole journey. I'm forever grateful.

This work was supported by the UKRI Centre for Doctoral Training in Artificial Intelligence & Music (EP/S022694/1).

"Art is not a reflection of reality, it is the reality of a reflection."
– Jean-Luc Godard, Filmmaker (Youngblood, 1998, pg. 29)

Table of Contents

1 Introduction 21

1.1 Research Gap 24

1.2 Case Study Domain: AI-based Music Composition 25

1.3 Research Questions 26

1.4 Methodological Approach 26

1.5 Contributions 27

1.6 Associated Publications 29

1.6.1 Supplementary Publications 29

1.7 Structure 30

I Designing A Questionnaire on Reflection in CSTs 33

2 State-of-the-Art: Reflection and CSTs 34

2.1 The Creative Process 34

2.1.1 Reflection 37

2.1.2 Engagement and Flow 43

2.1.3 Comparing Reflection and Engagement 45

2.2 Creativity Support Tools 46

2.2.1 Creativity Support Tools for Reflection 48

2.3 Evaluating Creativity Support Tools 52

2.3.1 Study Design Types and Settings 53

2.3.2 Data Collection: Questionnaires 57

2.3.3 Data Collection: Interviews and Video-Cued Recall . . 60

2.3.4 Data Collection: First-person Perspectives 61

2.4 Summary of Literature Critiques 62

3 Design of RiCE Version 1 65

3.1 Method Overview 65
3.2 Item Development 66
3.2.1 Preliminary Work 66
3.2.2 The Creative Professionals' Review 68
3.3 Scale Development 72
3.3.1 Participants 72
3.3.2 Procedure and Data Collection 73
3.3.3 Data Cleaning 74
3.3.4 Data Analysis Method 74
3.3.5 Results 75
3.4 Discussion 78
3.4.1 RiCEv1's Factors 80
3.4.2 Limitations 81
3.5 Conclusion 81

4 Evaluation of RiCE Version 1 82

4.1 Method 82
4.1.1 Participants 83
4.1.2 User Interfaces 83
4.1.3 Procedure and Data Collection 87
4.1.4 Data Analysis Method 88
4.2 Results 90
4.2.1 Confirmatory Factor Analysis 90
4.2.2 Test-Retest Reliability 91
4.2.3 Differentiation by Known Groups 92
4.2.4 Comparison with Existing Scales 92
4.3 Discussion 93
4.3.1 Limitations 95
4.4 Conclusion 95

5 Design of RiCE Version 2 96

5.1 Method 96
5.1.1 Dataset 97
5.1.2 Data Analysis Method 97
5.2 Design Iterations 98
5.2.1 Extending RiCEv1 98
5.2.2 Training Set: Iteration 1 98

5.2.3 Training Set: Iteration 2 102
5.2.4 Training Set: Iteration 3 102
5.2.5 Test Set: Iteration 4 and RiCEv2 Final Design 102
5.3 Discussion 102
5.4 Conclusion 104

II Reflection in AI-based Music Composition Tools 105

6 State-of-the-Art: AI-based Music Composition 106
6.1 Music Composition Process 107
6.2 Generative AI Architectures for Music 109
6.3 AI-based Creativity Support Tools 113
6.3.1 AI-based CSTs for Reflection 114
6.3.2 AI-based CSTs for Music Composition 115
6.4 Summary of Literature Critiques 119

7 Artist-Researchers' First-person Reflections in AI-based Music Composition 121
7.1 Method 121
7.1.1 Participants 122
7.1.2 AI Tools 124
7.1.3 Procedure 124
7.1.4 Data Collection 126
7.1.5 Data Analysis Method 129
7.2 Findings 130
7.2.1 Summary of First-person Accounts 130
7.2.2 Thematic Analysis Findings 136
7.3 Discussion 139
7.3.1 Limitations 141
7.4 Conclusion 142

8 RiCEv2 Analysis of Artist-Researchers' AI-based Music Composition 143
8.1 Data Analysis Method 143
8.2 Findings 144
8.2.1 Visualisations 144

8.2.2 Comparison of RiCEv2 Scores 147
8.3 Discussion 147
8.3.1 Novel Patterns 148
8.3.2 Patterns Similar to Current Research 149
8.3.3 Limitations 150
8.4 Conclusion 151

III Designing a New AI Music Composition Tool for Reflection 152

9 Iterative Design of wAIve 153
9.1 Method 154
9.1.1 Design Goals 154
9.1.2 Participants 154
9.1.3 Procedure 156
9.1.4 Data Analysis Method 157
9.2 Design Iterations 157
9.2.1 Initial Design 157
9.2.2 Iterations 1 to 4 (Data Scientists) 160
9.2.3 Iterations 5 to 8 (AI Musicians) 161
9.2.4 Iterations 9 to 12 (Designers) 164
9.2.5 Final Prototype 166
9.3 Discussion 168
9.3.1 Iterative Design Process 169
9.3.2 WAIve and its Features 170
9.4 Conclusion 172

10 Technical Implementation of wAIve 173
10.1 Overview 173
10.2 AI Music Generation 174
10.2.1 MusicVAE Generations 174
10.2.2 Metric-based Music Clustering 176
10.3 Animations 178
10.3.1 Flying Block Animations 178
10.3.2 Flashing Play Buttons 178
10.3.3 Curve Animations 179

10.3.4 Note Colour Coding and Fading 180
10.4 Summary 180

11 Evaluation of wAIve 181
11.1 Method 182
11.1.1 Participants 182
11.1.2 Procedure 185
11.1.3 Data Collection 186
11.1.4 Data Analysis Method 188
11.2 Findings 189
11.2.1 Descriptive Statistics Findings 189
11.2.2 Regression Analysis Findings 191
11.2.3 Thematic Analysis Findings 194
11.3 Discussion 198
11.3.1 Interplay between Reflection and Engagement 199
11.3.2 Reflection and AI Generated Content 201
11.3.3 Limitations 204
11.4 Conclusion 206

12 Comparison of RiCEv2 Results 207
12.1 Participants and Study Settings 207
12.2 Reflection 208
12.3 Reflection-on-Process 212
12.4 Reflection-through-Experimentation 213
12.5 Reflection-on-Self 215
12.6 Summary 217

IV Conclusions 218

13 Conclusions 219
13.1 Summary of Thesis and Main Findings 221
13.2 Answers to Research Questions 223
13.3 Recommendations for AI-based Musical CST Users 226
13.4 Limitations of the Methodological Approach 227
13.4.1 Open-ended Tasks and Settings 228
13.4.2 Study Tools 229

13.4.3 Quantitative: Questionnaires 229

13.4.4 Quantitative: Questionnaire Analysis 230

13.4.5 Qualitative: Interviews and First-Person Accounts . . 231

13.4.6 Qualitative: Thematic Analysis 232

13.5 Future Work 233

References **235**

Appendix **263**

List of Figures

2.1 Visualisations of models of the creative process. 35

2.2 Visualisations of models of the reflection process. 38

2.3 Relationship between challenge, skill and flow states. Too much challenge leads to anxiety, whereas too little challenge leads to boredom. Image based on Csíkszentmihályi (1990) and Nash (2011, pg. 81). 44

2.4 Custom views for artefacts in Kaleidoscope. The artefacts are organised to show development over time by default (left), but can be reorganised with flexible layouts (right). Layout histories can also be viewed (centre). Image from Sterman et al. (2023), CC BY 4.0. 50

2.5 Interface used to review and reflect on writing processes in 'Watch me write' (Carrera & Lee, 2022). Edits are shown in the timeline at the bottom of the user interface. Image from Carrera and Lee (2022), CC BY 4.0. 51

3.1 Participants' countries of origin in the RiCEv1 design study. . 73

3.2 SRIS scores in the RiCEv1 design study. 76

4.1 Participants' countries of origin in the RiCEv1 evaluation. . . 84

4.2 Interface for story-sentiment-visualiser. Participants were tasked with writing a positive story for two minutes. 85

4.3 Interface for sound-sketcher. Participants were tasked with writing a music composition for two minutes. 86

5.1 Iteration 1's training set correlations and alpha values for RiCEv2's design. 100

5.2 Iteration 2's training set correlations and alpha values for RiCEv2's design. 100
5.3 Iteration 3's training set correlations and alpha values for RiCEv2's design. 101
5.4 Iteration 4's test set correlations and alpha values for RiCEv2's design. 101

6.1 Visualisation of the variational auto-encoder architecture. Image adapted from Bryan-Kinns et al. (2021). 112
6.2 Interactive Generative Music Environment. Image from Hunt et al. (2020), CC BY 4.0. 116
6.3 Cococo's interface, where melodies can be augmented with AI generated harmonies. The sliders (right) can be used to steer the AI generated music output. Image from Louie et al. (2020), CC BY 4.0. 117
6.4 Laetitia Sonami's Spring Spyre. The metal springs stretched within the circular frame are mapped to audio output using interactive machine learning (Fails & Olsen, 2003). Image from Fiebrink and Sonami (2020), CC BY 4.0. 118

7.1 AI tools integrated into the composers' practices. **(A)** Real-time Audio Variational auto-Encoder (RAVE) (Caillon & Esling, 2021); **(B)** Neural resonator (Diaz, 2024). **(C)** Cue-Free Express + Pedal (CFEP) (Worrall et al., 2022); **(D)** Mark of Markov (MoM); **(E)** ProgGp (Loth et al., 2023); **(F)** Tidal-Fuzz (Wilson et al., 2021). 125
7.2 An example reflection board from the last session completed by P7. 128

8.1 Artist-researchers' reflection-on-process scores for each session of their composing. 145
8.2 Artist-researchers' reflection-on-self scores from RiCEv2 for each session of their composing. 145
8.3 Artist-researchers' reflection-through-experimentation scores from RiCEv2 for each session of their composing. 146

8.4 Model of the trade-off relationship between reflection-on-process and reflection-on-self. People selected ideas whilst listening in real-time and arranged their ideas after curation. 149

9.1 Overview of the iterative design process procedure for wAIve's design study. The researcher is orange. The participants are blue. 156

9.2 WAIve's initial design. 159

9.3 Note colours added to represent ascending and descending melodies in iteration 2 of wAIve's design. 160

9.4 Storyboard of the fading note colours added in iteration 4 of wAIve's design. 161

9.5 Iteration 4 of wAIve with an additional workspace and play buttons added to each block. 162

9.6 Iteration 6 of wAIve with multiple timelines. There is a different colour per timeline. 163

9.7 Flashing play buttons. Implemented in iterations 7 (left) and 8 (right) of wAIve's design. 164

9.8 Animated curves. These appear between AI-generated blocks (right) and blocks in a timeline (left) when new generations are added. 165

9.9 Iteration 10 of wAIve, with the dark colour scheme guided by the designers. 166

9.10 WAIve's final design. The features are labelled in yellow text. The text is not part of the user interface. WAIve can be tested online at https://codetta.codes/wAIve/. 167

10.1 Flowchart visualising the drawing loop which triggers AI generations and animations in wAIve. 174

10.2 Flowchart visualising the sequence for generating music blocks from MusicVAE. 175

10.3 Flowchart of the algorithm used to identify whether blocks should be flashed or not. 179

11.1 Goldsmiths MSI scores in the wAIve evaluation study. 184

11.2 SRIS scores in the wAIve evaluation study. 184

11.3 RiCEv2 scores in the wAIve evaluation study. 190

11.4 UEQ scores in the wAIve evaluation study. 190
11.5 Model 1 visualisation of RiCEv2 ∼ UEQ factors. 192
11.6 Model 2 visualisation of RiCEv2-Ex ∼ UEQ aesthetic. 192
11.7 Model 3 visualisation of RiCEv2-Cp ∼ UEQ factors. 193
11.8 Model 4 visualisation of RiCEv2-Se ∼ UEQ reward. 193
11.9 Model of interplay between reflection and engagement. Types of engagement are in blue. Types of reflection are in orange. . 199
11.10Model extending Figure 11.9 to show the influence of AIGC on reflection and engagement. Types of engagement are in blue. Types of reflection are in orange. 202

12.1 Goldsmiths MSI scores for CSTs across chapters. 209
12.2 SRIS scores for CSTs across chapters. 209
12.3 RiCEv2 scores for CSTs across chapters. 210
12.4 RiCEv2-Cp (Reflection-on-process) scores for CSTs across chapters. 212
12.5 RiCEv2-Ex (Reflection-through-experimentation) scores for CSTs across chapters. 214
12.6 RiCEv2-Se (Reflection-on-self) scores for CSTs across chapters.216

List of Tables

2.1 Summary of the main types of reflection. 42

2.2 Summary of study settings and data collection techniques for CST evaluations. 54

3.1 Three examples of items initially generated for the RiCEv1 design. See the full list in the appendix. 67

3.2 Creative professionals' backgrounds for the RiCEv1 design. Biographies are summarised from verbatim biographies found in the appendix. All participants were instructed to write their biographies to only include information that they consent to be published, as approved by the Queen Mary University of London's ethics committee. 69

3.3 Count of creative professionals scoring "Agree", "Netural" or "Disagree" for select items, sorted by the number of "Agree" scores. Items where 7 out of 10 or more creative professionals rated "Agree" were taken forward to the scale development phase. (R) denotes that the item's answer given by a participant in a user study should be reversed for analysis. 71

3.4 Counts of participants' choice of creative technology in the RiCEv1 design. 75

3.5 Loadings for the items in the RiCEv1 design. Values > 0.4 are in bold. 77

3.6 Summary of Chapter 3's main findings. 78

3.7 Items and instructions for administering and scoring RiCEv1. Cronbach's (1951) alpha values are also reported. 79

4.1 Participants' descriptive statistics for the RiCEv1 evaluation. 84

4.2 Fit metrics for RiCEv1's confirmatory factor analysis across conditions and interfaces. Acceptable metrics are in bold. . . 91
4.3 Intra-class correlations between the test and re-test measures for RiCEv1 and its factors. Significant measures in bold. . . . 91
4.4 Wilcoxon signed-rank tests across the interfaces for RiCEv1 on both test and re-test. Significant results are in bold. 92
4.5 Summary of Chapter 4's main findings. 93

5.1 The extensions to the RiCEv1 questionnaire. These are later refined to create RiCEv2. Question numbering is from the original question list for RiCEv1's design in Table 3.3. 99
5.2 Items and instructions for administering and scoring RiCEv2. 103

7.1 Overview of the artist-researchers' characteristics. 123
7.2 Summary of the artist-researchers' chosen AI tools. Labelling (a) to (f) refers to the image labelling in Figure 7.1. 125
7.3 Interview questions used alongside the reflection board process in the artist-researcher user study. 129
7.4 Summary of Chapter 7's main findings. 139

8.1 RiCEv2 scores calculated from Chapter 3's data (top) and for this chapter's study (bottom). The DAWs include Cubase (n=2), Garageband (n=2), Ableton (n=2), Logic and FL Studio. 147
8.2 Summary of Chapter 8's main findings. 148

9.1 Goldsmiths MSI and SRIS scores for wAIve's design study. Participants are labelled by their expertise where DS = data scientists, AIM = AI and music interfaces, and DE = design. 155
9.2 Interview questions used as the basis for the semi-structured interviews in wAIve's design study. Questions marked with † were added as iterations progressed. 158
9.3 Summary of Chapter 9's main findings. 168

11.1 Participants' demographics and descriptive statistics for the wAIve evaluation study. 183
11.2 Interview questions used alongside the video-cued recall process in the wAIve evaluation study. Questions marked with $ were also used for wAIve's design. 187

11.3 Comparisons between the UEQ total score metric and RiCEv2 total score metric in the wAIve evaluation study. 191
11.4 Significant regression models for the wAIve evaluation study. Aspects of engagement predict aspects of reflection. 191
11.5 Summary of Chapter 11's main findings. 198

12.1 Summary of Chapter 12's main findings. 217

13.1 Summary of thesis's main findings. 220
13.2 Summary of data collection and analysis methods used across the thesis. 227

List of Abbreviations

ACM	Association for Computing Machinery
ACM CHI	ACM Conference on Computer-Human Interaction
AI	Artificial Intelligence
AIGC	Artificial Intelligence Generated Content
CSI	Creativity Support Index
CST	Creativity Support Tool
ACM C&C	ACM Conference on Creativity and Cognition
DAW	Digital Audio Workstation
GPT	General Purpose Transformer
HCI	Human-Computer Interaction
MIDI	Musical Instrument Digital Interface
ML	Machine Learning
MSI	Musical Sophistication Index
NIME	New Instruments for Musical Expression
RNN	Recurrent Neural Network
RiCE	Reflection in Creative Experience Questionnaire
SRIS	Self Reflection and Insight Scale
TA	Thematic Analysis
UEQ	User Engagement Questionnaire
UX	User eXperience
VAE	Variational Auto-Encoder
VCR	Video-Cued Recall

Chapter 1

Introduction

Creativity and technology have together influenced culture and the arts. Rosetta Tharpe pushed guitars into distortion to revolutionise rock and roll (Wald, 2023). Visual effects artists at Industrial Light and Magic pioneered novel techniques for overlaying images, used in technology such as Photoshop (Manovich, 2011), to drive innovations in film and media.

Despite its importance to culture, research on creativity only emerged at pace in the 1950s, within the field of psychology (Guilford, 1950). The trajectory of psychology research on creativity has been broadly described in three waves (Frich et al., 2019). The first wave describes the initial rise in the 1950s after Guilford's (1950) address to the American Psychological Association. Guilford (1950) described creativity as a core component of their structure of intellect model, characterising creativity as the number of divergent thoughts a participant can invent for an object. The second wave critiqued this characterisation of creativity from a social-constructivist viewpoint. It argued that the quantification of divergent thoughts ignored the social aspects of creativity (Amabile, 1983; Csíkszentmihályi, 1999). The third wave, in its emergence (Frich et al., 2019), focuses on collaborative and digital creativity.

The multifaceted and subjective nature of creativity means that it is ill-defined. Colton and Wiggins (2012) describe creativity as an essentially contested term; its correct definition requires continuous debate on its definition. Creativity has thus been studied in a variety of ways, characterised in psychology by Rhodes's (1961) four Ps of creativity, described below.

Person: Creativity can be viewed as a personality trait. Indeed, creativity was seen as an innate characteristic of geniuses, leading to studies of their personality (Albert & Runco, 1999).

Product: Creativity can be examined in creative products, under the assumption that creativity is inherent to its outcomes. Distinctions are proposed between creative outputs that are historically important and judged by people as creative post-hoc (H-Creative), or perceived by an individual as creative for themselves (P-creative) (Boden, 1991). For example, P-creative outcomes occur in everyday activities such as report writing or cooking (Richards, 2010).

Process: Creativity can be studied by examining *how* people create. This contrasts with research on creative products, focusing instead on how said product came to be. For example, Locher (2010) examined the process of visual artists from archives of interim sketches, video observations and sensor data. De Bono's (1985) thinking hats, a seminal work on brainstorming, identifies a set of mindsets that people can apply in the creative process to think from different perspectives.

Press: Creativity does not occur in a vacuum. Thus, there are studies on the wider culture and social milieu where creativity happens. Csíkszentmihályi (1999) proposed a systems view of creativity where creativity is socially constructed as the result of interactions between people and their cultural environment. Amabile (1983) emphasised the social nature of creativity, arguing its definition was innate amongst people within a creative domain.

Despite the diverse range of approaches to creativity, a common definition of creativity cited in psychology is: where *novel* work is produced which is *useful* for a group at some point in time (Stein, 1953). However, there is debate around *novelty* and *usefulness* and their relevance to creativity. For example, both the arts and the sciences use creativity. Yet, whilst the arts emphasise novelty and individual self-expression, the sciences emphasise finding practical solutions and problem solving (Glăveanu & Kaufman, 2021, pg. 18).

Boden (1991) similarly defines creativity as the ability to come up with ideas which are *novel* (to the person who has the idea, cf. P-creativity) and *valuable*. However, they add that novel ideas lead to different types of surprise. They introduce three types of processes that lead to novel and valuable ideas, each with different types of surprise attached:

Combinational Creativity: where ideas that are not typically similar are brought together, for example, an analogy in poetry. This leads to surprises where something unusual occurs and you are surprised it happened.

Exploratory Creativity: where new ideas are developed which fit within the conceptual space of a previously accepted style. This leads to the surprise of seeing something atypical within a given style.

Transformational Creativity: where new ideas change the rules of a culture. For example, cubism brought a new perspective to art. This leads to the surprise of seeing something that was thought impossible, and can take a long time to be accepted.

Hewett et al.'s (2005) taxonomy captures the range of characterisations of creativity described above. Its factors are used to situate creativity-related research within the broader range of perspectives on creativity. The taxonomy includes the factors of:

- product/persons/**process**;
- **personal**/social;
- **domain-specific**/general;
- and **individual**/collaborative.

This thesis focuses on the aspects of creativity highlighted in bold font above. It focuses on individuals' subjective experiences of their creative process. It is situated within the Human-Computer Interaction (HCI) field and advances a research gap in the current state of Creativity Support Tool (CST) evaluation, outlined further below.

1.1 Research Gap

Since the early 2000s (Fischer, 2004; Shneiderman, 2002), the field of CSTs has investigated the design and use of technology to support people in the creative process (Frich et al., 2019). This includes for a variety of creative contexts: from music (Bryan-Kinns & Hamilton, 2012) to fashion (Jeon et al., 2021) to dance (Fdili Alaoui, 2019). This variety of CSTs means a multitude of theories are used for their design and evaluation. These theories relate to various experiential aspects of the creative process (Remy et al., 2020).

Engagement, where people are drawn in and immersed in an activity when using a computer (Bryan-Kinns et al., 2007; Chapman, 1997; Wu, 2018), has dominated the design and evaluation of CSTs. For example, engagement is central to seminal guidelines for CST design (Resnick et al., 2005; Shneiderman et al., 2006). Engagement is also prominent in the Creativity Support Index (Cherry & Latulipe, 2014), which has enabled the systematic evaluation of CSTs over the last decade.

Reflection is another crucial aspect of the creative process (Candy, 2019). It is central to the informal, experiential learning which occurs in creative experiences with technology (Candy, 2019, pg. 178), where people build tacit knowledge by cyclically interacting with a CST and reflecting on its outcomes (Schön, 1983). However, reflection contrasts engagement when characterised by moments of self-awareness and introspection. For example, during moments of reflection, people often step back, disrupting their attention during moments of engagement (Moon, 2013; Sharples, 1996; Wilson et al., 2023). Reflection also requires effortful thinking (Kahneman, 2011), contrasting with the autotelic and fun-based interactions that occur in moments of engagement (O'Brien & Toms, 2008).

Few CST evaluations address reflection, and even fewer have systematically assessed reflection as part of a mixed-methods user study, which is the standard for CST evaluation (Hewett et al., 2005). Cox et al. (2025) show that of 173 user studies on CSTs in the last 10 years, only 6 (including research published from this thesis) applied measures of self-reflection. This demonstrates that there are limited frameworks and no consensus on how to operationalise reflection for CST evaluation.

This thesis argues that a systematic evaluation of reflection in CST interaction is needed to characterise and support the user's creative process. The focus on reflection challenges and advances on the dominance of engagement-based evaluations in the CST field.

1.2 Case Study Domain: AI-based Music Composition

This thesis selects Artificial Intelligence (AI)-based music composition as its case study domain to focus the investigation of reflection in CST interaction to a specific creative context.

Music composition is fundamental to human evolution and social development (Freeman, 1998), and a fundamental element of intangible cultural heritage (Unesco, 2020). As with other creative domains, music composition presents an open-ended challenge where technology mediates the potential creative possibilities (Magnusson, 2019). Moreover, computer music is an active area of research, exemplified by various publication venues such as the New Instruments for Musical Expression conference (Poupyrev et al., 2001) or the International Society for Musical Information Retrieval conference (Downie et al., 2009). These qualities justify music composition as a rich and specific area in which to investigate reflection in CST interaction.

Furthermore, the world is currently experiencing a new wave of AI research (Xu, 2019). Whilst AI algorithms for generating music autonomously have been developed for decades (Harley, 2002; Hiller & Isaacson, 1957), recent breakthrough techniques in AI have led to systems capable of convincingly imitating existing music (Carnovalini & Rodà, 2020). This thesis focuses on these generative AI systems, following the definition of generative AI as: models built from datasets to produce new data with similar statistical properties (ibid). The outputs from these AI systems are referred to as AI-Generated Content (AIGC).

The elevated interest in AIGC has renewed previous discussions (Cornock & Edmonds, 1973; Lubart, 2005) on the opportunities for creativity support using computers. For music composition, generative AI has been applied to CSTs to extend the possibilities of how people interact with music. For example, generative AI can complete menial composition tasks up to acting similar to a collaborative musical partner (Jourdan & Caramiaux, 2023).

Within the creative industries, interest in AIGC has thus grown significantly (Caramiaux et al., 2019). There have also been calls for more human-centred AI research (Garibay et al., 2023; Shneiderman, 2022), including in music (Jourdan & Caramiaux, 2023). This justifies the case study of *AI-based* music composition as a timely case study domain. It represents the state-of-the-art in CST design and resonates with calls for human-centred AI.

1.3 Research Questions

In line with the thesis's argument and case study domain, the following research questions are addressed:

- **RQ1:** How is reflection characterised in people's open-ended interaction with AI-based CSTs designed for music composition?
- **RQ2:** What is the interplay between characterisations of reflection and engagement in people's open-ended interaction with an AI-based CST for music composition?

These questions enable an advance in the state-of-the-art by providing new characterisations of reflection for users of musical AI-based CSTs (RQ1) and positions the findings in contrast to existing engagement-focused tools (RQ2). The answers to the research questions are in Section 13.2.

1.4 Methodological Approach

This thesis applies a mixed methods approach to address the research questions above. This follows the standard for evaluating CSTs (Hewett et al., 2005). Each user study involves people interacting with a CST and giving post hoc feedback. The studies are conducted systematically, limiting how long people use a CST and guiding them through a study protocol. The study tasks are open-ended rather than comparative – for example, there are no comparisons between a system with and without AI. The open-ended approach is used to investigate typical characteristics of creative processes (Kerne et al., 2013), in line with the thesis research questions.

1.5 Contributions

This thesis makes two central claims of new knowledge.

Contribution 1: There is new knowledge on how reflection is characterised in people's music composition with AIGC.

The state-of-the-art shows little research on how reflection is characterised in CST interaction, and fewer studies using questionnaires to systematically characterise reflection in mixed-methods evaluations. For the thesis's case study domain of AI-based music composition, there are no studies that systematically characterise how people reflect when using AIGC. This thesis advances on the state-of-the-art by presenting systematic characterisations of people's open-ended interaction with AI-based musical CSTs. It characterises reflection in AI-based musical CST interaction as a balance of reflection-on-process, reflection-through-experimentation, and reflection-on-self. Reflection-on-process is shown as common in CST interaction, whilst reflection-on-self is less common – particularly in music interaction contexts. Reflection-through-experimentation occurs early in the AI-based music composition process, and is common when users are unfamiliar and learning about an AI tool. These characterisations are based on findings from the study chapters described below.

Chapter 8 characterises that artist-musicians reflect-on-process when curating AIGC in real-time, and self-reflect when organising their curated AIGC. The value is that users of AI-based musical CSTs can adapt their practice to invoke more of each type of reflection (see Section 13.3). For example, users should spend more time focused on organising AIGC to encourage more self-reflection.

Chapter 11 characterises the interplay between reflection and engagement for computer science students in AI-based music composition. It uses a novel AI-based CST as a case study to identify common reflection patterns. The interplay between reflection and engagement is that moments of focused attention occur alongside engagement, and that self-reflection occurs alongside people's feelings of a rewarding experience. It also shows interplay between reflection and focused attention, and the conditions for AI tools under which this happens: users reflect in moments of focused attention when uninterrupted or when learning new interactions from an AI. It is the first study to identify relationships between reflection and engagement based on a systematic CST evaluation, in the AI-based music composition context. The value to AI-based musical CST users is that these novel characterisations inform recommendations for when and how users should use AIGC to invoke different types of reflection (see Section 13.3).

Contribution 2: There is a new contribution of knowledge in the form of a novel questionnaire for reflection in CST interaction.

The state-of-the-art for CST evaluation uses questionnaires based on the engagement theory of flow (Csíkszentmihályi, 1990). There are no questionnaires focused on reflection in CST interaction. This thesis contributes a new questionnaire enabling the systematic assessment of reflection, thus providing a methodological advance on existing CST evaluation tools. Developed across Chapters 3, 4 and 5, the new questionnaire characterises reflection as three conceptually meaningful factors: reflection-on-current-process, reflection-on-self and reflection-through-experimentation. Its suitability is demonstrated by its use in Chapters 8, 11 and 12. Indeed, its findings underpin the characterisations of reflection contributed in Chapters 8 and 11. The value is that other CST researchers can use the questionnaire to systematically assess different types of reflection in different tools and study conditions.

The contributions above relate to the study of CSTs within the field of HCI. The research is appropriate for publication at the ACM Creativity and Cognition conference (Candy & Edmonds, 1999). This contrasts music-led areas such as New Instruments for Musical Expression (Poupyrev et al., 2001) and AI-focused areas such as Computational Creativity (Colton & Wiggins, 2012). This thesis does not contribute to the development of novel AI systems, which would be a more typical contribution to the Computational Creativity field. Although this thesis touches on these related areas, the terminology and literature used throughout are selected to reflect the focus on CST and HCI research.

1.6 Associated Publications

Portions of the work detailed in this thesis have been presented in peer-reviewed, international scholarly publications, as listed below.

Chapters 7 and 8 are based on the following publication:

- **Ford, C.**, Noel-Hirst, A., Cardinale, S., Loth, J., Sarmento, P., Wilson, E., Wolstanholme, L., Worrall., K., & Bryan-Kinns, N. (2024) Reflection-Across AI-based Music Composition. *Proceedings of the 2024 ACM Conference on Creativity and Cognition.* https://doi.org/10.1145/3635636.3656185 **Note: Honourable Mention Award.**

Chapters 3 and 4 are based on the following publication:

- **Ford, C.** & Bryan-Kinns, N. (2023) Towards a Reflection in Creative Experience Questionnaire. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3544548.3581077

Section 2.3 is based on portions of the following publication written by the author:

- Lerch, A., Arthur, C., Bryan-Kinns, N., **Ford, C.**, Sun, Q. and Vinay, A. (2025) Survey on the Evaluation of Generative Models in Music. arXiv preprint arXiv:2506.05104. **Note: Accepted to ACM Computing Surveys. In publication process.**

1.6.1 Supplementary Publications

The following publications, whilst not directly included in this thesis, showcase research completed by the author during their PhD with relevance to HCI, AI and music.

- Saitis, C., Del Sette, B.M., Shier, J., Tian, H., Zheng, S., Skach, S., Reed, C. N., **Ford, C.** (2024) Timbre Tools: Ethnographic Perspectives on Timbre and Sonic Cultures in Hackathon Designs *Proceedings of the 2024 ACM International Audio Mostly Conference - Explorations in Sonic Cultures (AM '24)* https://doi.org/10.1145/3678299.3678322

- Bryan-Kinns, N., Noel-Hirst, A., & **Ford, C.** (2024) Using Incongruous Genres to Explore Music Making with AI Generated Content

Proceedings of the 2024 ACM Conference on Creativity and Cognition. https://doi.org/10.1145/3635636.3656198

- Bryan-Kinns, N., **Ford, C.**, Zheng, S., Kennedy, H., Chamberlain, A., Lewis, M., Hemment, D., Li, Z., Wu, Q., Xiao, L., Xia., G, Rezwana, J., Clemens, M., Vigliensoni, G. (2024) Explainable AI for the Arts 2 (XAIxArts2). *Proceedings of the 2024 Conference on Creativity and Cognition* https://doi.org/10.1145/3635636.3660763
- Bryan-Kinns, N., **Ford, C.**, Chamerlain, A., Benford, S., Kennedy, H., Li, Z., Wu, Q., Xia, G. & Rezwana, J. (2023) Explainable AI for the Arts: XAIxArts. *Proceedings of the 2023 Conference on Creativity and Cognition* https://doi.org/10.1145/3591196.3593517
- Bryan-Kinns, N., Banar, B., **Ford, C.**, Reed, C. N., Zhang, Y., Colton, S., and Armitage, J. (2021) Exploring XAI for the Arts: Explaining Latent Space in Generative Music. *Proceedings of the 1st Workshop on eXplainable AI Approaches for Debugging and Diagnosis at NeurIPS 2021* https://xai4debugging.github.io/files/papers/exploring_xai_for_the_arts_exp.pdf

1.7 Structure

The remaining chapters are structured in four parts.

Part I: Designing A Questionnaire on Reflection in CSTs

Chapter 2 surveys the state-of-the-art creativity and CST research relevant to the thesis. It shows that there is little CST research on reflection, and fewer studies using questionnaires to examine reflection in systematic mixed-methods evaluations. This identifies the need for a new questionnaire to systematically evaluate reflection in CST interaction.

Chapter 3 thus presents the development of the Reflection in Creative Experience Questionnaire Version 1 (RiCEv1). The chapter finds four conceptually meaningful factors (reflection-on-current-process, reflection-on-self, reflection-on-past-experience and reflection-through-experimentation).

Chapter 4 evaluates RiCEv1 and shows it has good content validity and can successfully differentiate between the different types of reflection that are common in different CSTs.

Chapter 5 further updates RiCEv1 to RiCEv2, including removal of the reflection-on-past-experience factor, to improve upon its construct validity for use in the thesis case study domain.

Part II: Reflection in AI-based Music Composition Tools

With a questionnaire for systematically assessing reflection in CST interaction in place, the thesis moves to its case study on AI-based music composition.

Chapter 6 surveys the state-of-the-art in AI and music composition. It shows that few studies discuss how people reflect using AIGC in CSTs, and none have assessed reflection in the AI-based music composition context. This justifies a mixed-methods evaluation of reflection in people's music composition with AI-based CSTs.

Chapter 7 thus presents first-person accounts of artist-researchers' reflections, each using a different AI music tool. RiCEv2 was applied after each hour of composing to examine their reflection over time.

Chapter 8 triangulates the first-person accounts with patterns in the artist-researchers' RiCEv2 responses, systematically characterising their reflection.

Part III: Designing a New AI Music Composition Tool for Reflection

The previous part characterised reflection patterns across a variety of AI-based music composition practices. This justifies the need for a new tool with reflection support as a central design goal, to focus the evaluation.

Chapter 9 thus documents the iterative design of a new AI-based music tool, named wAIve. Its design is based on feedback from participants paired by their skills in data science, AI and music interfaces, and design.

Chapter 10 describes the technical implementation details for the features identified as emphasising reflection in wAIve.

Chapter 11 presents a systematic mixed-methods evaluation of reflection in AI-based music composition, using wAIve as a case study tool. RiCEv2 is applied to evaluate wAIve amongst a sample of computer science students.

A characterisation of reflection is presented that demonstrates interplay between focused attention and reflection, and the conditions where this occurs when using AIGC.

Chapter 12 compares the wAIve RiCEv2 scores with the RiCEv2 scores calculated for other CSTs assessed in the thesis. Patterns in reflection are identified that are common amongst the various CSTs, including those specific to AI-based music composition.

Part IV: Conclusion

Chapter 13 answers the thesis's research questions. In line with the thesis argument, it outlines how the findings can be used to support AI-based musical CST users in their composition process. A reflection on the methodological approach and its limitations is then presented, followed by opportunities for future study to bring the thesis to a close.

Part I

Designing A Questionnaire on Reflection in CSTs

Chapter 2

State-of-the-Art: Reflection and CSTs

This chapter surveys the state-of-the-art creativity and the Creativity Support Tool research (CST) that is relevant to this thesis. It shows that there is little CST research on reflection, and fewer studies using questionnaires to systematically examine reflection as part of a mixed-methods evaluation. This justifies the design of the questionnaire in the following chapters.

This chapter begins with research on the creative process. Second, CSTs are introduced, tracing their origins from the seminal first workshop on CSTs (Shneiderman et al., 2006). Third, the chapter reviews the state-of-the-art for CST evaluation in Human-Computer Interaction (HCI). The experiential qualities of reflection and engagement are emphasised throughout. The chapter closes with a summary of the literature critiques.

2.1 The Creative Process

Models of the creative process have been developed in several disciplines, including design (Cross, 2008; Sanders & Stappers, 2008) and psychology (Amabile, 1996; Wallas, 1926). Five key processes are visualised in Figure 2.1 from across disciplines and described below. The processes are selected for their relevance to CST research. For example, the design process models are frequently used in CST research, given the overlap between HCI and design (Gaver, 2012).

Wallas (1926) developed one of the first models of the creative process, formalising ideas from people's introspective accounts (see Lubart (2001)). It

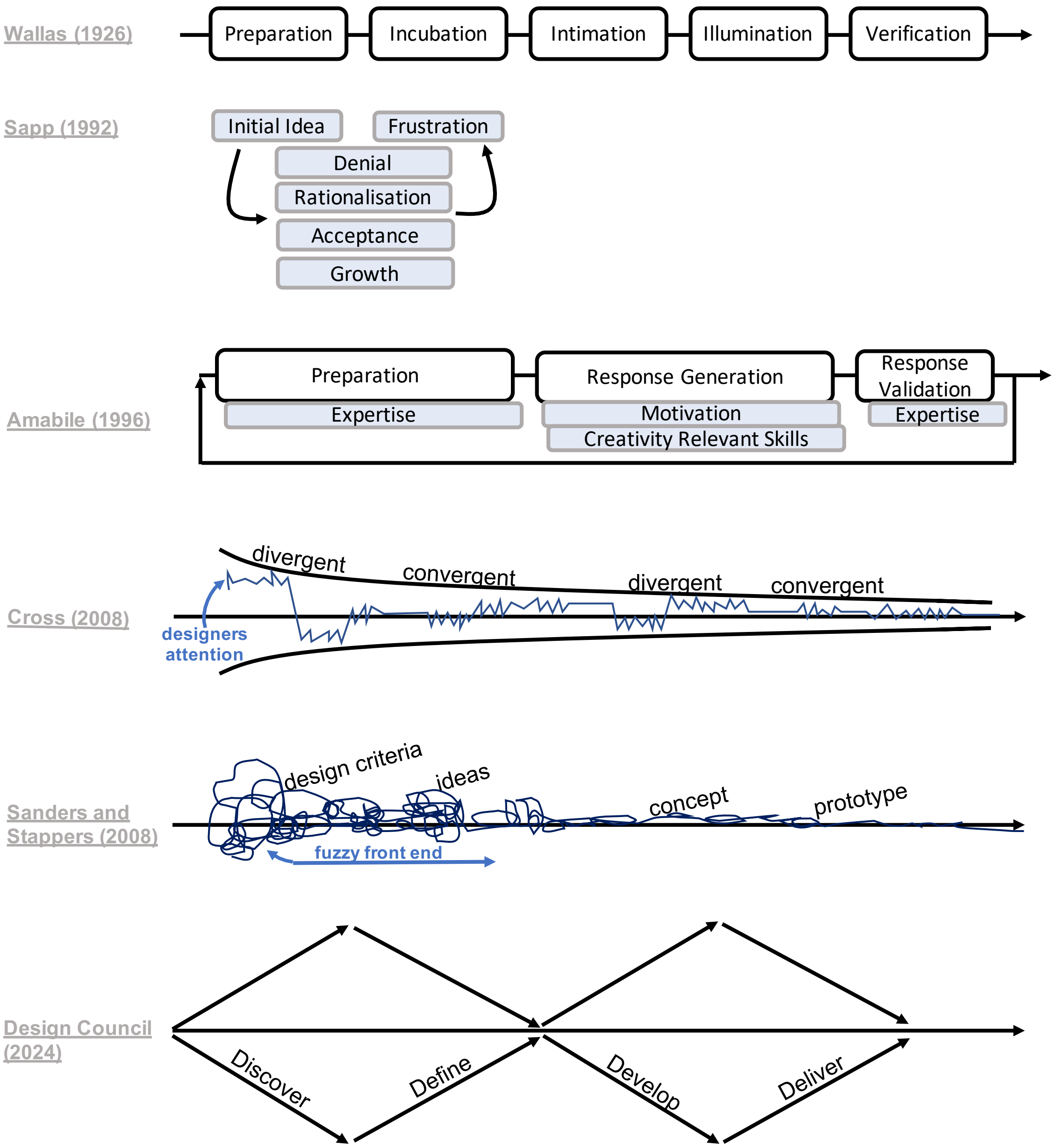


Figure 2.1: Visualisations of models of the creative process.

consists of four linear stages. First, *preparation*, where a person consciously collects ideas and plans their creation. Second, *incubation*, where the mind unconsciously processes potential problems. Third, *illumination*, where a promising idea breaks into consciousness. Fourth, *verification*, where the idea is consciously tested.

However, Wallas's (1926) model has several limitations. Thus, extensions to the model have been proposed. For instance, Sapp (1992) added a stage of frustration following the preparation phase, where a person moves from denial of creative limitations towards acceptance. Amabile (1983, 1996) critiqued that Wallas's (1926) stages are not linear (people move back and forth between them) and notes how creativity is amplified in the process by fostering motivation and expertise.

In the field of design, models of the creative process focus on divergent and convergent thinking. Divergence and convergence are also influential in seminal psychology research on creativity (see Guilford (1950)). Sanders and Stappers (2008) describe how designers start with a "fuzzy" initial idea and move between divergence and convergence before reaching a final design. The Design Council's double diamond[1] model presents the design process as alternating stages of convergence and divergence. These stages are *discover* (divergent thinking to identify ideas), *define* (converging on a definition), *develop* (divergent ways to develop the idea), and *deliver* (converging on the final output).

The creative process models above describe how creative work unfolds. However, even when models include iterative movement between stages (Amabile, 1996), there is often a more dynamic blend of stages and frequent overlap. The models also provide little insight into the experiential aspects of creative practice (Nelson & Rawlings, 2009). These aspects can include feelings of awareness, anxiety, and control (see Section 2.1.2). Feelings of failure (Hazzard et al., 2019; Kim et al., 2015) or surprise (Candy, 2019, pg. 67) also influence the creative experience. This thesis focuses on the experiential aspect of *reflection* and argues it is crucial in CST interaction contexts. Interdisciplinary literature on reflection is thus introduced below.

[1] https://www.designcouncil.org.uk/our-resources/the-double-diamond/

2.1.1 Reflection

Reflection is crucial to the creative process (Candy, 2019; Wiggins, 2006). For example, reflection is central to practice-based research across creative domains (Candy, 2019; Guillaumier, 2016). HCI studies also often report reflections on creative user experiences, including in sketching (Lewis et al., 2023), music (Sturm, 2022), and dance (Fdili Alaoui, 2019).

Reflection is ill-defined and interpreted differently across HCI research (Baumer, 2015; Baumer et al., 2014; Bentvelzen et al., 2022; Fleck & Fitzpatrick, 2010) and related fields, such as education (Boud et al., 1985; Habermas, 1987; Kolb, 1984; Roldan et al., 2021) and design (Dalsgaard & Halskov, 2012; Rivard & Faste, 2012). A common understanding is that reflection involves moments where people sit back in quiet contemplation (Moon, 2013; Wilson et al., 2023). Interdisciplinary literature describes reflection as having an outcome (Atkins & Murphy, 1993; Boud et al., 1985; Boyd & Fales, 1983; Steinaker & Bell, 1975) or as a process for clarifying uncertain situations (Dewey, 1933; Kolb, 1984; Schön, 1983). Moon (2013) offers a pragmatic definition of reflection as "a basic mental process with either a purpose or an outcome or both[...] applied in situations where material is ill-structured or uncertain and where there is no obvious solution". However, this definition is broad and incomplete, lacking insight into specific qualities of reflection and how it develops over time.

Fields such as education and nursing have proposed models for how reflection develops over time (Atkins & Murphy, 1993; Boud et al., 1985; Boyd & Fales, 1983; Dewey, 1933; King & Kitchener, 1994; Kolb, 1984; Steinaker & Bell, 1975). Five models of the reflection process are shown in Figure 2.2. The models are selected as they represent the seminal ideas on the reflection process that have influenced subsequent works in CST and HCI research. For instance, Cho et al. (2022) drew upon models of reflection to summarise seven steps for reflection in craft-making: document, search, observe, organise, compare, connect and iterate.

The models in Figure 2.2 also emphasise how reflection emerges from experiential learning: the "construction of learning from observations [...] in some practical situation" (Moon, 2013, pg. 20). Kolb (1984) identifies that learning occurs when a person reflects on new experiences, forms generali-

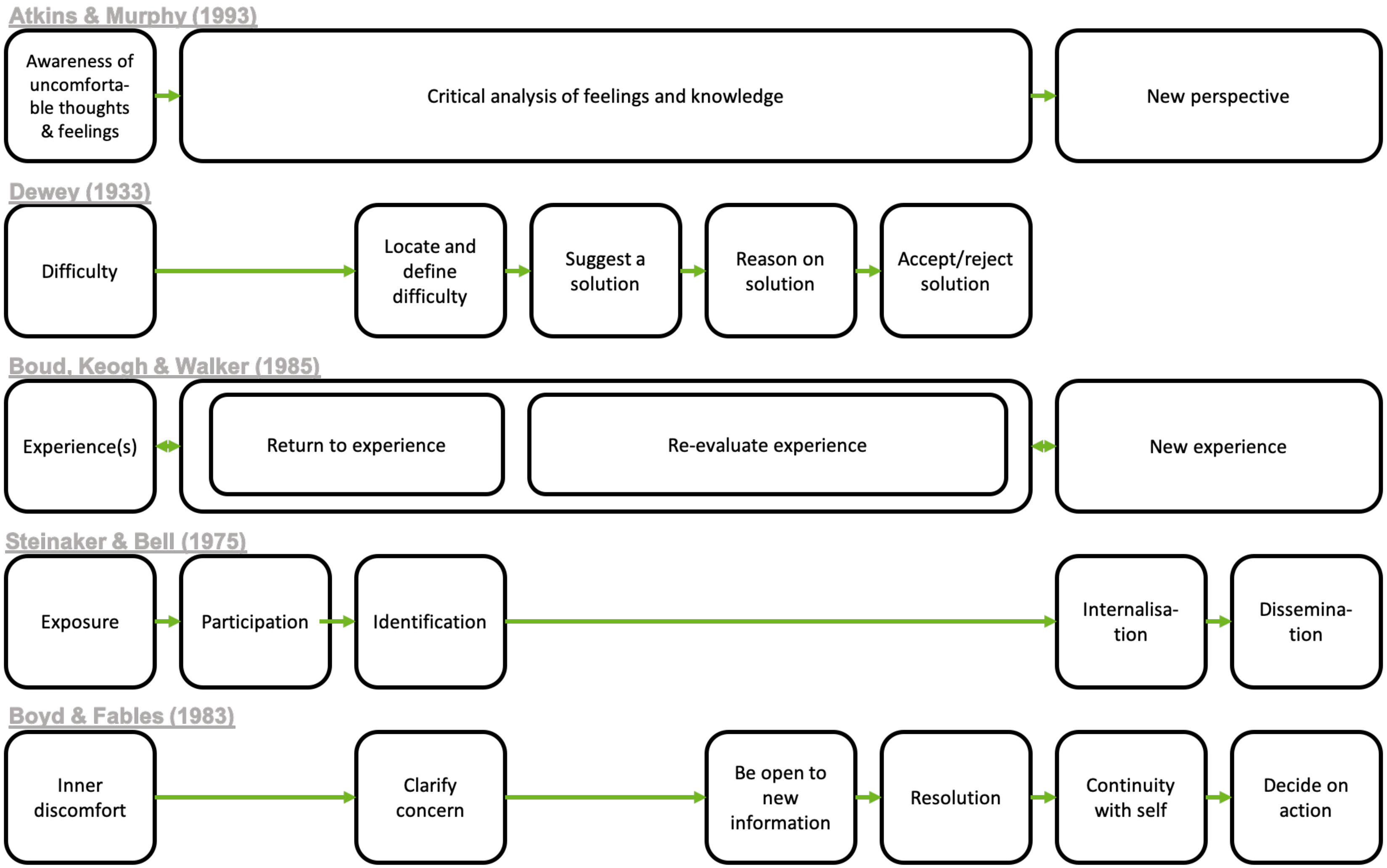


Figure 2.2: Visualisations of models of the reflection process.

sations, and tests these generalisations in new contexts. Kolb's description connects with Dewey (1933), who views reflection as an inquiry where ideas are formulated, considered, and either accepted or rejected. Boud et al. (1985) describes reflection as a process where people re-evaluate new experiences by considering their *feelings*, resulting in new perspectives or behaviour change. The critique of these models is that they focus on how people look back on an experience to reflect post-hoc, not moments of reflection that occur *during* an experience.

Schön (1983) developed two characterisations of reflection that have been influential across disciplines, including HCI (Baumer, 2015; Baumer et al., 2014; Fleck & Fitzpatrick, 2010). They introduce the concepts of *reflection-in-action* (when a person's behaviour does not result in the expected outcome, so they reflect on their actions in-the-moment to solve the issue) and *reflection-on-action* (reflecting after or away from an activity). With the distinction of reflection-in-action, Schön emphasised tacit ways of knowing which occur in creative disciplines, where knowledge is built mainly from practice-based learning. In contrast to the models of reflection described above, Schön's reflection-in-action notably emphasises reflection *during* an experience. This challenged the dominant learning techniques used in universities at the time, which emphasised rote recall instead of practitioner-centred approaches (Schön, 1987).

Schön's ideas resonate with Kahneman's (2011) theory of fast and slow thinking. Fast thinking is instantaneous, intuitive, and emotional, while slow thinking is deliberate and requires effort. In HCI, Norman (1993) offers a similar distinction of experiential cognition (fast reactions without effort or delay) and reflective cognition (a slower consideration of ideas). Whilst Kahneman refers to thinking instead of reflection, the term reflection is used here to imply moments of thinking which lead to positive changes or an outcome (Moon, 2013).

2.1.1.1 Aspects of Reflection

This thesis proposes the following aspects of reflection as relevant to creative user experiences. The aspects are based on the research above on the reflection process (see Figure 2.2) and HCI perspectives on reflection (see Section 2.2.1).

Breakdown: Reflection can occur in moments of breakdown (Baumer, 2015): where a person's actions map to outcomes against their intuition (Dewey, 1933; Kolb, 1984; Schön, 1983). This is shown in the initial stages of models in Figure 2.2, including Atkins and Murphy's (1993) uncomfortable stage, Dewey's (1933) difficulty stage, and Boyd and Fables's (1983) discomfort stage.

Comparison: When reflecting, people think back on previous experiences (Boud et al., 1985; Kolb, 1984; Schön, 1983) or evaluate previous actions in new contexts (Norman, 1993). This relates to Boud et al.'s (1985) stages in Figure 2.2 of returning to and re-evaluating an experience. People also compare themselves to others when reflecting (Bentvelzen et al., 2022).

Impact: When reflecting, people consider the broader implications of their actions. This includes how their actions influence different people and cultures (Fleck & Fitzpatrick, 2010).

Inquiry: Baumer's (2015) and Dewey's (1933) models (see Figure 2.2) show that people generate, test and revise hypotheses iteratively whilst reflecting.

Motivation: For reflection to occur, just having the tools is insufficient. People must also *decide* to engage in reflection (Fleck & Fitzpatrick, 2010; Grant et al., 2002; Slovák et al., 2017).

Openness: People remain open to new experiences (Kolb, 1984) and paths of inquiry (Boud et al., 1985; Boyd & Fales, 1983) when reflecting, acknowledging that variables can change. This is shown as a distinct phase of Boyd and Fables's (1983) model (see Figure 2.2).

Transformation: People change their understandings (Atkins & Murphy, 1993; Boud et al., 1985; Boyd & Fales, 1983) and question assumptions when reflecting (Baumer, 2015; Kolb, 1984). For example, Steinaker and Bell's (1975) model in Figure 2.2 presents internalisation of new understandings as a distinct phase. Atkins and Murphy (1993) and Boud et al. (1985) also emphasise contemplating new perspectives and experiences as distinct phases.

Trustworthiness: Norman (1993) describes that people sometimes contemplate the trustworthiness of different information when reflecting. Fleck and Fitzpatrick (2010) and Dewey (1933) describe how information which is most pragmatic or which corroborates the most perspectives is selected. Untrustworthy information can lead to breakdowns or difficulty (cf. Dewey (1933) in Figure 2.2) or feelings of inner discomfort (cf. Boyd and Fales (1983) in Figure 2.2).

Candy (2019) contributes the only characterisations of reflection explicitly for creative contexts. These characterisations are:

- **Reflection-for-action:** Reflecting on the possible actions to take in preparation for creating.
- **Reflection-in-the-making-moment:** Reflecting on decisions during interaction with materials.
- **Reflection-at-a-distance:** Taking an objective step back to evaluate one's art.
- **Reflection-on-surprise:** Reflecting on unexpected occurrences.

However, Candy's characterisations are derived from qualitative interviews with creative practitioners and require operationalisation for application in user studies. This means they can not be used to systematically evaluate and compare reflection in different creative experiences. This contrasts with the characterisations of reflection in questionnaires, which produce comparable, quantitative scores directly from users (see Section 2.3.2).

Candy also lists **non-reflective actions** as a type of reflection, where creative practitioners act more on their emotional, intuitive thinking. This is similar to Kahneman's (2011) fast thinking in that non-reflective actions occur instantaneously in lieu of considering additional factors. Non-reflective actions also resonate with theories on engagement, such as flow theory (Csíkszentmihályi, 1990), which have dominated CST evaluation. Theories on engagement are thus introduced in the following section, then compared and contrasted with reflection.

To summarise, Table 2.1 provides a summary of the main different types of reflection, based on the aspects proposed above and the types of reflection in the literature.

Table 2.1: Summary of the main types of reflection.

Reflection Type	Definition	Citation
Reflection-in-action/in-the-making-moment	Reflection during an experience. This can be triggered by breakdowns (when outcomes differ from intuition) or moments of discomfort. It involves iterative inquiries where people adjust their actions, and relates to fast modes of thinking as decisions are often based on intuitions (Kahneman, 2011; Norman, 1993).	(Schön, 1983) (Candy, 2019) (Dewey, 1933) (Baumer, 2015) (Atkins & Murphy, 1993)
Reflection-on-action	Reflection after an experience such as using a technology. It involves making comparisons with past actions or to others, and assessing next steps, cf. transformation. This engages slower modes of thinking (Kahneman, 2011; Norman, 1993) to contemplate and assess next steps.	(Schön, 1983) (Bentvelzen et al., 2022) (Boud et al., 1985) (Kolb, 1984)
Reflection-at-a-distance	Reflection after stepping away from a work, for a more objective perspective.	(Candy, 2019) (Boud et al., 1985) (Kolb, 1984)
Reflection-for-action	Reflection on possible action to take in preparation before creating.	(Candy, 2019)
Transformative reflection	Reflection where the re-evaluation of an experience or interaction leads to changed understanding, behaviour, or perspective.	(Boud et al., 1985) (Atkins & Murphy, 1993) (Boyd & Fales, 1983) (Baumer, 2015) (Slovák et al., 2017) (Fleck & Fitzpatrick, 2010)

2.1.2 Engagement and Flow

Engagement is a complex, multifaceted concept, interpreted differently across disciplines (Doherty & Doherty, 2018). Some view engagement as a person's state with different levels. For example, Edmonds et al. (2006) proposes distinctions between *passive engagement*, where people momentarily notice a system, and *active engagement*, where people return to and develop skills whilst using a system. They further describe three categories of engagement in interactive art systems: *attractors* (sparking initial interest), *sustainers* (holding participation during an initial encounter) and *relaters* (holding an ongoing relationship). Bilda et al. (2008) describe phases of engagement over time with an interactive artwork, from *adaptation* (initial understanding of how a system behaves) to *learning* (developing a mental model of the system), up to *deeper understanding* (evaluating an artwork at a conceptual level).

Engagement can also represent moments of connection between people (Sidner et al., 2004), such as where they creatively spark together (Bryan-Kinns & Hamilton, 2012; Bryan-Kinns et al., 2007). Furthermore, engagement relates to Costello and Edmonds's (2007) categories of pleasure, which include: *creation* (pleasure from making), *discovery* (pleasure from finding new directions) and *difficulty* (pleasure from overcoming challenges).

In HCI, engagement is understood as moments when people are drawn in and attentive during interaction with a computer (Bryan-Kinns et al., 2007; Chapman, 1997; Wu, 2018). O'Brien et al. (2018) proposes four aspects of reflection in HCI:

- **Focused attention:** Moments of immersion or concentration with a system.
- **Aesthetic qualities:** The visual appeal of a system.
- **Perceived usability:** How easy a system is to use.
- **Reward:** How fulfilling using a system is.

Engagement is closely related to the mental state of flow, where people feel in control and lose self-awareness (Csíkszentmihályi, 1990). In the field of positive psychology, where flow theory originated, flow states represent an optimal state of happiness (Nakamura & Csíkszentmihályi, 2009; Seligman

& Csíkszentmihályi, 2014). Flow states are also often described as conducive to creativity, in part due to their basis in interviews with creative experts such as artists or musicians (Csíkszentmihályi, 1990)

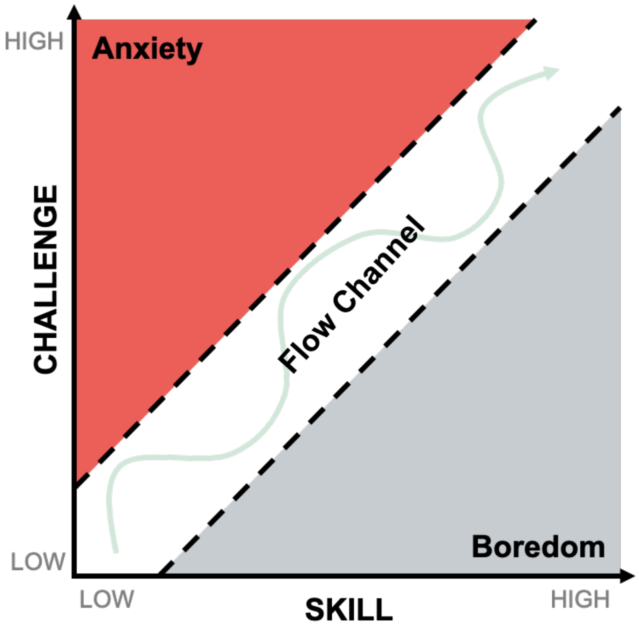


Figure 2.3: Relationship between challenge, skill and flow states. Too much challenge leads to anxiety, whereas too little challenge leads to boredom. Image based on Csíkszentmihályi (1990) and Nash (2011, pg. 81).

Flow states occur when challenge is met with ability, as visualised in Figure 2.3. Too much challenge and people experience anxiety. Too little challenge leads to boredom. Increasing challenge over time as people's skills develop supports intrinsic motivation as a flow state is maintained. Recall that Amabile (1982) describes intrinsic motivation as conducive to creativity (see Section 2.1).

The theory of flow also posits nine aspects of a flow state, listed below based on Nakamura and Csíkszentmihályi (2009):

1. Clear goals and feedback.
2. A challenging activity that requires skill.
3. The merging of action and awareness.
4. Concentration on the task at hand.

5. A sense of control.
6. The transformation of time.
7. The loss of self-consciousness.
8. The sense of time becomes distorted.
9. The autotelic experience.

These aspects of flow states coincide with aspects of engagement. For example, moments of focused attention (O'Brien et al., 2018) are similar to the immersion experienced in a flow state. There are also distinctions between aspects of a flow state and aspects of engagement. For example, a condition for a flow state is that it is intrinsically motivated and requires long-term focus. In contrast, engagement is extrinsically motivated and can occur while multitasking (O'Brien & Toms, 2008). Other HCI researchers see engagement as a subset of flow, for example, where people can have less control or unclear goals (Doherty & Doherty, 2018). Furthermore, engagement has been operationalised as an observable construct, whereas flow states are characterised by subjective feelings which cannot be observed (Nakamura & Csíkszentmihályi, 2009; Seligman & Csíkszentmihályi, 2014).

2.1.3 Comparing Reflection and Engagement

During moments of reflection, people often step away from an activity, disrupting their attention during moments of engagement (Moon, 2013; Sharples, 1996; Wilson et al., 2023). People also reflect on themselves and the world around them (Boud et al., 1985; Boyd & Fales, 1983; Kolb, 1984), contrasting the immersive aspects of flow states where self-awareness fades away (Sheldon et al., 2015; Tang & Zhou, 2020). Reflection also requires effortful thinking (Kahneman, 2011) in comparison to the more automatic, reactive interactions that occur during moments of engagement (O'Brien & Toms, 2008). Whereas the theory of flow emphasises the need for clear goals, in reflection, people tend to navigate ambiguous (Gaver et al., 2003) and uncertain goals.

Creativity-related literature has also compared reflection and engagement. Candy (2019, pg. 62) introduces a perspective challenging flow theory to the field of creativity support, with the caveat that it was derived from goal-oriented activities such as sports and not creative domains. This perspective is Montero's (2016) cognition-in-action principle. The principle of cognition-in-action states that people are at their optimal performance when consciously reflecting and applying mental processes. This contrasts with aspects of flow states, such as forgoing self-awareness and introspection. Sharples (1996) describes a relationship between reflection and engagement in the context of creative writing. They describe how writers move back and forth between periods of *reflection* (reviewing material, contemplating and planning ideas) and *engagement* (producing material following the reflective phase). Thus, a push-and-pull relationship between reflection and engagement occurs in creative processes.

2.2 Creativity Support Tools

The previous section reviewed literature on the creative process, with focus on the experiential aspects central to this thesis: reflection and engagement. This section reviews how the creative process has been supported with technology. It introduces the state-of-the-art in CST design, including CSTs that support reflection. It shows that CSTs have primarily been designed to support engagement, whilst comparatively few focus on reflection, despite its importance.

The design and evaluation of technology that supports people's creativity have been studied in HCI since the early 2000s (Frich et al., 2019). To date, the HCI sub-field of creativity support investigates CST design across different domains and user types: from children drawing on an iPad to professional artists and designers (Frich et al., 2019). For example, CST research is published frequently in dedicated communities such as the ACM Conference on Creativity and Cognition (Candy & Edmonds, 1999).

Frich et al. (2019) reviewed 143 CST papers from the ACM Digital Library to develop the following definition of a CST:

> "A Creativity Support Tool runs on one or more digital systems, encompasses one or more creativity-focused features, and is employed to positively influence users of varying expertise in one or more distinct phases of the creative process" (Frich et al., 2019, pg. 10).

This definition is broad, encompassing many subsets of tools. For example, in the music domain, this definition would include software applications, plugins, and digital musical instruments.

Early CST research was grounded in psychology theories, which described creativity as a socially constructed phenomenon. Shneiderman (2002) proposed a systems view of creativity where computers make “people more creative more often” by supporting the social activities of: connecting (to previous work), relating (to mentors and peers), creating (to explore different possible solutions), and donating (to disseminate results to others). Fischer (2004) also proposed to support creativity with computers from a social perspective, identifying social barriers computers could address, such as removing the need for travel with video conferencing technology.

Taking inspiration from Shneiderman (2002) and Fischer (2004), the seminal first workshop on CSTs brought together 25 research leaders and graduate students to discuss CST design and evaluation (Shneiderman et al., 2006). As the first international event focusing on CSTs, its reports became highly influential contributions. In particular, design principles derived from discussions at the workshop (Resnick et al., 2005) have informed the design of many CSTs (Frich et al., 2019). The design principles include:

Support exploration: CSTs should allow for easy exploration of alternative designs and different aspects of a design. This helps users take new directions that are unclear from the start of a project. For instance, using an undo feature helps users feel more comfortable testing possibilities.

Low threshold, high ceiling, and wide walls: CSTs should be easy to start using (low threshold), but powerful and sophisticated (high ceiling), so that people can create a wide range of possibilities. Wide walls refer to designing CSTs to support a range of creative projects. This principle connects with others from the seminal first workshop on CSTs, such as to make a CST as simple as possible (low threshold) and to support different interaction styles (wide walls).

Support collaboration and support open interchange: In line with Fischer (2004) and Shneiderman (2002), CSTs should support sharing ideas to foster collaboration and social interaction.

The CST design principles were informed by the theory of flow (Csíkszentmihályi, 1990) (see Section 2.1.2). As Shneiderman et al. (2006) highlights in his summary of the seminal first workshop on CSTs, the assumption is that CSTs which nurture a flow state will better support creativity. For example, the principle of supporting a low threshold and high ceiling aligns with the concept of increasing challenge in flow theory. The CST design principle to make tools as simple as possible also mirrors balancing challenge and ability in flow theory; a more straightforward interface lowers challenge and helps users enter a flow state.

The influential literature above demonstrates how CST design has traditionally focused on engagement and flow. There are fewer CST designs for reflection, despite its importance in creative user experiences. The historical background on reflection in HCI and the few existing CSTs that support reflection are introduced below.

2.2.1 Creativity Support Tools for Reflection

HCI research on reflection emerged near 2010 (Baumer, 2015). This was catalysed by a CHI[2] workshop (Sas & Dix, 2009) and two key review papers (Baumer et al., 2014; Fleck & Fitzpatrick, 2010).

Fleck and Fitzpatrick (2010) first synthesised interdisciplinary literature on reflection to design a pragmatic framework for interaction designers. They outlined how technology could support increasingly sophisticated levels of reflection, from reflecting on previous actions (for example, by visualising log data) to reflecting on broader social and cultural ramifications (for example, by visualising data people had not considered). Building on this literature review, Baumer (2015, 2014) created a framework of three dimensions that support the design of technology for reflection. The three dimensions were: *breakdown* (puzzling moments which spark reflection), *inquiry* (hypothesis testing and exploration of old ideas) and *transformation* (leading to changes in thinking and understanding).

Slovák et al. (2017) highlighted an issue with the design recommendations for reflection presented above. They argued that new and different visuali-

[2] The ACM Conference on Human Factors in Computing Systems (CHI) is the top conference in the field of HCI.

sations of data would not necessarily spark reflection without sufficient user motivation. They thus considered ways designers could scaffold the reflection process. For example, designers could focus on creating emotional and interpersonal experiences to motivate people's reflection.

Bentvelzen et al. (2022) presented concrete suggestions for supporting reflection in design by extending Baumer's (2014) review. Unlike previous reviews of reflection, they considered 98 interactive systems designed to enhance reflection from the ACM Digital Library and the Apple App Store. They found several common design features tied to aspects of reflection. These included allowing users to revisit their data (to prompt introspection) and to share data on social media (to encourage comparison and conversation). However, their review focused on personal informatics applications: systems with the functional goal to help people make decisions based on their personal data (Baumer et al., 2014; Bentvelzen et al., 2022; Dijk et al., 2017). This contrasts with the open-ended goals of CST interactions.

Other areas of HCI have emphasised the importance of reflection, for example, within design processes (Sengers et al., 2005) or in people's reactions to ambiguous artefacts (Gaver et al., 2003). Reflection is also important in Slow Technology (Hallnäs & Redström, 2001), which encourages people to be more mindful and less reactive when using technology. Furthermore, HCI often discusses reflection as part of metacognition, which supports learning in environments such as computer science classrooms (Li, Zhang, & O'Rourke, 2024; Li, Chen, et al., 2024). However, little research has focused on reflection in the CST domain. Examples of CSTs which support reflection are discussed below.

Documentation for reflection: To support reflection, CSTs can help people to document their creative processes. ReflectionSpace (Sharmin & Bailey, 2013) and Kaleidoscope (Sterman et al., 2023) let users archive artefacts and notes, which can be visualised in different time scales. Figure 2.4 shows some of Kaleidoscope's various options, allowing saving and reloading of different layouts for reflection-on-action (Schön, 1983). Dalsgaard and Halskov (2012) also present a CST for researchers to document their work over time, providing opportunity for reflection on collections of material with text-based memos. Mosaic (Kim et al., 2017) is a CST that presents a social media platform for people to document their creative process to encourage reflection amongst a community of practitioners.

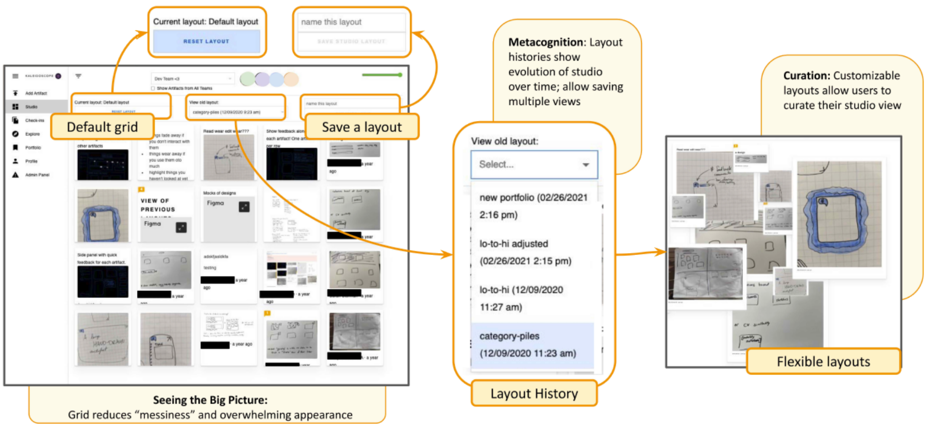
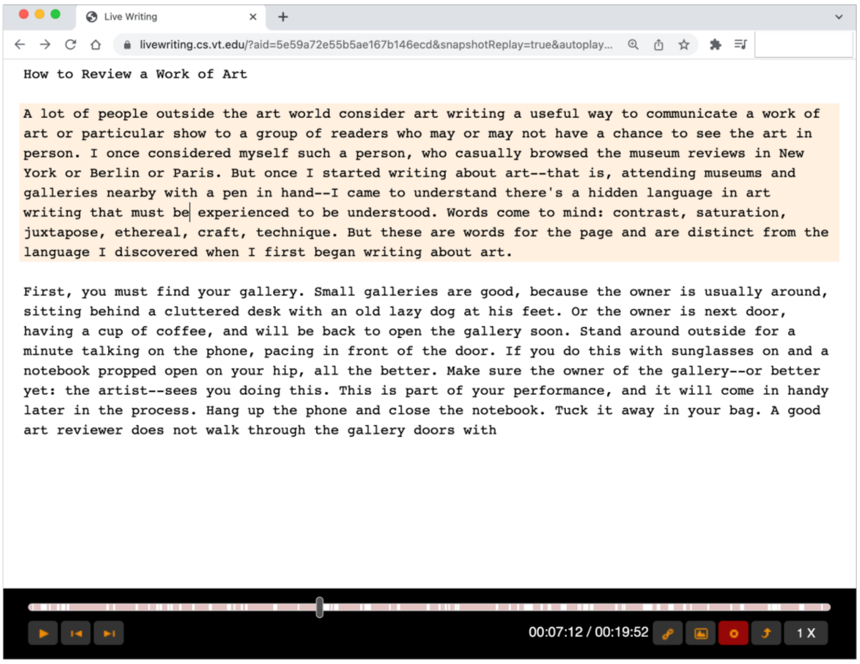

Current layout: Default layout
RESET LAYOUT
name this layout
SAVE STUDIO LAYOUT
Default grid
Save a layout
Metacognition: Layout histories show evolution of studio over time; allow saving multiple views
View old layout:
Select...
new portfolio (02/26/2021 2:16 pm)
lo-to-hi adjusted (02/26/2021 2:15 pm)
lo-to-hi (12/09/2020 11:27 am)
category-piles (12/09/2020 11:23 am)
Layout History
Curation: Customizable layouts allow users to curate their studio view
Flexible layouts
Seeing the Big Picture:
Grid reduces "messiness" and overwhelming appearance

Figure 2.4: Custom views for artefacts in Kaleidoscope. The artefacts are organised to show development over time by default (left), but can be reorganised with flexible layouts (right). Layout histories can also be viewed (centre). Image from Sterman et al. (2023), CC BY 4.0.

The critique is that documentation tools only support reflection-on-action (Schön, 1983) because they capture the creative process for post-hoc analysis. This contrasts with reflections which occur *during* interaction with a CST. Furthermore, documenting work only provides the opportunity for reflection – users must be motivated to engage in looking back (Slovák et al., 2017).

Replaying Creative Process: CSTs have been designed that allow people to replay their creative process. 'Watch me write' (Carrera & Lee, 2022), shown in Figure 2.5, introduced a tool where writers can playback their own or others' writing sessions. Writers found that real-time playback was engaging, bringing awareness to and sparking self-reflection on their writing style. Spin (Tseng & Resnick, 2016) takes the form of a turntable and camera, which captures animations of how projects evolve over time. The time to produce an animation was fast enough to support engagement, yet long enough to leave space for reflection. Surprises in the animations also sparked engagement. However, similar to documentation tools, these CSTs focus on prompting reflection after a creative activity, not during a creative experience.



Live Writing
livewriting.cs.vt.edu/?aid=5e59a72e55b5ae167b146ecd&snapshotReplay=true&autoplay...

How to Review a Work of Art

A lot of people outside the art world consider art writing a useful way to communicate a work of art or particular show to a group of readers who may or may not have a chance to see the art in person. I once considered myself such a person, who casually browsed the museum reviews in New York or Berlin or Paris. But once I started writing about art--that is, attending museums and galleries nearby with a pen in hand--I came to understand there's a hidden language in art writing that must be experienced to be understood. Words come to mind: contrast, saturation, juxtapose, ethereal, craft, technique. But these are words for the page and are distinct from the language I discovered when I first began writing about art.

First, you must find your gallery. Small galleries are good, because the owner is usually around, sitting behind a cluttered desk with an old lazy dog at his feet. Or the owner is next door, having a cup of coffee, and will be back to open the gallery soon. Stand around outside for a minute talking on the phone, pacing in front of the door. If you do this with sunglasses on and a notebook propped open on your hip, all the better. Make sure the owner of the gallery--or better yet: the artist--sees you doing this. This is part of your performance, and it will come in handy later in the process. Hang up the phone and close the notebook. Tuck it away in your bag. A good art reviewer does not walk through the gallery doors with

Figure 2.5: Interface used to review and reflect on writing processes in 'Watch me write' (Carrera & Lee, 2022). Edits are shown in the timeline at the bottom of the user interface. Image from Carrera and Lee (2022), CC BY 4.0.

Offering new perspectives: There are examples of CSTs which present different perspectives on information to encourage reflection. Cho et al. (2022) present an interface for creating wire-based jewellery through motion. The interface visualises a database of jewellery designs clustered by their similarity. The clustering encourages comparison and offers various views of the jewellery for reflection. DramatVis (Hoque et al., 2022) offers word cloud and timeline visualisations for qualities of different characters in creative writing, to support reflection on different perspectives. SonAmi (Belakova & Mackay, 2021) supports writers' reflection by reading back their work through a speaker attached to a coaster whenever they sip coffee.

However, whilst the metric-based visualisations and audio representations used in these CSTs can support reflection on specific aspects, they often make little functional sense in relation to broader creative contexts or intentions. For example, whether a character's name appears more often than another's does not necessarily reflect their intended impact on a story. Kaschub and

Smith (2009) and Whittall (2011) also argue that (musical) intentions are often indescribable and based on intuition.

In summary, there are three key critiques of CSTs for reflection. First, most CSTs support reflection-on-action (Schön, 1983) and not reflection *during* their use. Second, features such as unique visualisations of data lack functional correspondence to people's creative intentions, which often cannot be articulated (Kaschub & Smith, 2009; Whittall, 2011). Third, users require motivation to engage in reflection with these tools (Slovák et al., 2017).

It is not possible to systematically inspect the different types of reflection that occurred in people's interaction with the CSTs above. This is due to a variety of methods being used to assess reflection. This motivates the need for an operationalisation of reflection which enables its systematic evaluation in various CST designs. To understand how this can be achieved, the state-of-the-art methods and techniques for CST evaluation are introduced in the following section, focusing on reflection and engagement.

2.3 Evaluating Creativity Support Tools

CST evaluations fall within the intersection of creative practice and computer science (Hewett et al., 2005), and draw from interdisciplinary fields encompassing HCI such as psychology (Preece et al., 2011, pg. 10). This poses tension between epistemological perspectives on research design; for example, HCI techniques borne from psychology strive to identify generalisable models of how people use technology, whilst arts and humanities-inspired approaches offer insight into the subjectivity of how artists interpret their technology use (Candy & Edmonds, 2018).

Hewett et al. (2005) outlined the standard for CST evaluation based on discussions at the seminal first workshop on CSTs. They advocated for a mixed-methods approach, where quantitative and qualitative data collection approaches are triangulated to allow for corroboration between findings. Hewett et al.'s (2005) research methods have been widely used in subsequent HCI research on media and arts technology (Bryan-Kinns & Reed, 2023). These methods also share aims with human-centred AI research (Garibay et al., 2023; Shneiderman, 2022), which advocates for evaluating AI systems from a person-first perspective.

To understand the state-of-the-art in mixed-method CST evaluation, HCI study design types and their settings are described below. Then, different data collection approaches frequently used in mixed-methods CST evaluations are outlined. An overview of the study design types and data collection methods, and the assumptions for their use are shown in Table 2.2.

2.3.1 Study Design Types and Settings

Early HCI research focused on functional aspects of user interaction, such as a system's usability (Nielsen, 1994). Later waves of HCI focused more on the subjective user experience (Bødker, 2015). Approaches to structuring studies for usability and user experience are broadly split into two types: *comparative* and *open-ended* experiments.

Comparative Studies: In a comparative study, two or more versions of an interface are tested to allow comparison between versions – often referred to as A-B testing (Nielsen, 1994, pg. 178). For instance, researchers could test a CST with and without AI to see if the AI affects different measures. For example, Louie et al. (2020) tested two versions of a system, one with and one without AI-steering tools, to evaluate the effect of the AI-steering tools (see Section 6.3.2). Here, the tool without AI features would be the *baseline*. Different groups could also be compared to test the effect of their characteristics on the same interface.

Comparative experiments are understood as more objective and show the impact between groups or between interface designs. However, the findings are often limited to the manipulated features. Thus, there is less opportunity to capture spontaneous and serendipitous interactions, which are common in creative contexts. Comparative studies are also less suited to research goals intended to characterise aspects of the broader user experience; people's thoughts and feelings are focused on the dependent variable instead of the fuller range of frequent interaction patterns that contribute to the user experience aspect being investigated.

Open-ended Studies: Instead of comparing conditions directly, an open-ended study examines people's interaction and subjective feelings when using an interface or technology. They typically involve more open tasks than con-

Table 2.2: **Summary of study settings and data collection techniques for CST evaluations.**

Approach	**Assumptions**	**Captures**	**Use in Thesis**
Comparative	Variables can be controlled across CST designs. Typically viewed as objective and conducted in lab settings. Low ecological validity.	Effects between groups or interface designs.	Ch. 4
Open-ended	Common features of interaction are observable in one condition. Typically viewed as more subjective than comparative studies, but more objective than in-the-wild studies. Allows for use of open-ended tasks to improve realism. However, use of lab settings limits realism.	Subjective qualities and features of an interaction.	Ch. 7, 8, 9, 11
Online	Data at a global scale helps identify generalisable findings. Can yield noisy data. No control over environment, nor opportunity to correct misinterpretations.	Diverse users' assessments of an interaction.	Ch. 3, 4
In-the-wild	Real-world use and context is most important. All confounding variables seen as features of the study. Most ecologically valid.	Rich data representing real-world use and contextual behaviour.	Ch. 7, 8
Questionnaires	Self-reported measures reflect users' experience, attitudes, and perception. Understood as standardised and generalisable.	Scores from users on their perception of their experience and attitudes.	**Use in Thesis**
	Tools	**Captures**	
	Usability (Bangor et al., 2008; Blackwell & Green, 2000; Hart, 2006)	Perceived complexity, cognitive effort, ease of use.	N/A
	User Experience (Laugwitz et al., 2008)	A broad range of user experience aspects, such as stimulation, novelty, and satisfaction.	N/A
	Creativity (Cherry & Latulipe, 2014)	A tools capacity to support a range of creativity-related factors.	Ch. 4
	Engagement (Jackson et al., 2008; O'Brien et al., 2018)	Measures of engagement-related factors such as focused attention and aesthetic appeal.	Ch. 11
	Reflection (Bentvelzen et al., 2021; Grant et al., 2002)	People's capacity for self-reflection or their reflection on personal data. **There are no CST targeted tools.**	Ch. 3, 4, 7, 8, 9, 11
Interviews	Can be structured, semi-structured or open-ended. Assumes that questions do not lead people to certain responses and that their recall of an experience is accurate.	Thoughts, motivations, and user reasoning.	**Use in Thesis**
	Tools	**Captures**	
	Think-aloud	Verbalised thoughts during use. Understood as having limited realism.	N/A
	Video-Cued Recall	Reflections on past activity. Understood as more realistic than think-aloud.	Ch. 11
First Person	Users' insights provide richer data than interviews. Relies on writing skill. Limited generalisability.	Deep subjective and personal insights.	Ch. 7

trolled studies. For example, Tchemeube et al. (2023) tasked people with exploring a generative AI plugin and collected post-hoc feedback using questionnaires and interviews. The limitation of open-ended studies is that they tend to be more subjective in terms of task choice and study setting than comparative experiments. However, this suits investigations of experiential aspects of the user experience, where the focus is on understanding the features of users' data collected on their thoughts and feelings of an interaction, and the selected evaluation techniques.

There is also often overlap between open-ended and comparative studies to balance their benefits. Open-ended studies can, for example, take place in a controlled lab setting typically used in comparative studies, yet allow creative interaction to occur in a more natural manner. Indeed, controlled and open-ended studies are applicable to different settings, such as in controlled settings, online, and in their places of happening, described below.

2.3.1.1 Controlled Settings

Controlled experiments place participants in distraction-free environments such as a research lab (Nielsen, 1994, pg. 200). Controlled settings account for confounding variables as there are no outside environmental influences (Hewett et al., 2005). This makes lab settings well suited to comparative studies to isolate the variable under study, for example, a user interface design feature.

While isolating external variables in a lab is useful, this raises issues of ecological validity: how applicable a study's results are to real-world practice, where external factors are inevitable. Tasks can be more open-ended (for example, to write a piece of music) to support ecological validity, depending on the CST being tested and the goals of the evaluation. An example CST is Cococo (Louie et al., 2020), where participants were tasked with composing music for a fictional character from a game in a controlled setting.

2.3.1.2 Online

Online evaluation is well-suited to conduct HCI studies across a large sample size, including for creative work (Oppenlaender et al., 2020). In online settings, CSTs can be hosted online for user interaction or available as a download. Audiences across the globe can then be reached using survey platforms such as Amazon Turk[3] or Prolific[4], and filtered for their characteristics such as nationality or technical skills. For example, Ben-Tal et al. (2021) hosted an AI-based CST online to examine data logs of how people generated content and how they tweaked values to modify and curate outputs (see Section 6.3.2).

However, online settings can lead to unreliable data collection, such as when people complete surveys as quickly as possible to claim a monetary reward (Müller et al., 2014). People are also far more likely to misinterpret online tests, and researchers are not directly present to correct misunderstandings (ibid.). Study designs requiring users to engage for an extended period, for example, to study aspects such as flow states (Csíkszentmihályi, 1990), can also become impractical and expensive. Similar to controlled environments, online environments are also often not where CSTs are used, undermining ecological validity. Furthermore, establishing the exact environment in which participants have completed online surveys is not always possible.

2.3.1.3 In-the-wild

Research-in-the-wild (Benford et al., 2013; Chamberlain et al., 2012; Crabtree et al., 2012) approaches best fit open-ended study designs where technology is studied in its real-world places of use. For example, ethnographic approaches place researchers within real-world scenarios to identify findings based on their observations, field notes and observed patterns of behaviour (Crabtree et al., 2012). This can include behaviours that people have naturally exhibited when using CSTs.

However, ethnographies must take place over long time periods for rich insights to occur. Strategies have been identified to capture rich data in shorter time periods, such as to focus on "key informants" (Millen, 2000). For example, in a 48-hour hackathon setting, Saitis et al. (2024) observed how developers and musicians collaborated to create digital tools focused on the musical quality of timbre. Many confounding factors also influence 'in-the-wild' settings; thus, the effect of different variables cannot be isolated.

An example of a study with generative music where ethnographic-inspired observations were collected is Fiebrink et al. (2012). Fiebrink et al. (2012)

[3] https://www.mturk.com/

[4] https://www.prolific.com/

"recorded text minutes of the group's activities, discussion topics, and specific questions, problem reports, and feature requests" (pg. 137) for composers using their Wekinator (Fiebrink et al., 2011) system in weekly seminars. Sterman et al.'s (2023) approach to evaluating Kaleidoscope in its actual place of use, the HCI classroom, can also be described as ethnographically-inspired. They used various data sources collected over time, including interviews, surveys, meetings, usage data, and students' written reflections.

In summary, open-ended study designs are most appropriate for studies focused on experiential aspects such as reflection. Comparative studies are more suitable for evaluating distinct aspects of CST designs. The choice of research setting between controlled and in-the-wild environments impacts the ecological validity of the study. However, a balance can be struck, for example, by using open-ended tasks in a controlled setting. With an understanding of typical CST study contexts, the following sections introduce data collecteharsenlengyeldion methods and techniques typically combined as part of a CST mixed-methods evaluation.

2.3.2 Data Collection: Questionnaires

Questionnaires can quantify both subjective feelings of a system's usability (such as Bangor et al. (2008)) and more experiential aspects of a user experience (such as Laugwitz et al. (2008)). Typically, these questionnaires use a ordinal scale and ask how much a user agrees with a given statement, for example, from 1–5. This operationalisation enables systematic comparison between questionnaire scores. Questionnaires also avoid interpretation from researchers (such as with qualitative methods) as users provide scores directly. It is best practice to use established questionnaire measures to connect with research that has also used the scale or shown its reliability (Müller et al., 2014). However, established questionnaires often use statements that do not directly fit the study context, which can inflate the length of a user study.

Several questionnaire measures are commonly used to evaluate the usability of technology, such as the NASA Task Load Index (Hart, 2006) and the Standard System Usability Scale (Bangor et al., 2008). The Cognitive Dimensions of Notations questionnaire (Blackwell & Green, 2000) assesses cognitive qualities of a user interface, such as whether it has many hidden elements, and it has been used in several CST evaluations (Bellingham, 2022; Hunt et al., 2020; Nash & Blackwell, 2012). However, these usability-focused measures do not capture experiential qualities such as reflection and engagement.

User experience-oriented questionnaires include the User Experience Questionnaire (Laugwitz et al., 2008), which focuses on several aspects, including stimulation and novelty. The wide range of factors means the scale captures a breadth of data on the user experience; however, this can distract from more focused study goals and lead to longer study completion times. There are a limited number of established questionnaires focused on specific experiential aspects. This is in part because developing questionnaires is a skill that requires researchers to expend time and effort in developing non-trivial skills in survey creation, which is itself a field of study (Boateng et al., 2018).

2.3.2.1 Measuring Engagement

For engagement and its related theories, such as flow theory (Csíkszentmihályi, 1990), many questionnaires have been applied to CST evaluations (Cox et al., 2025). For example, the Flow-Short Scale (Jackson et al., 2008) has been used to measure retrospective assessments of people's feelings of a flow state. For example, Nash and Blackwell (2012) adapted and used the Flow-Short Scale questionnaire to identify correlations between factors of flow states and music interaction. As the Flow-Short Scale focuses on people's feelings of flow, using it alongside other measures of interaction was necessary to be able to interpret its scoring in the context of a CST evaluation.

Engagement was highly influential in the design of the Creativity Support Index (CSI; Cherry and Latulipe, 2014), which assesses the capacity of a tool to support creativity. The CSI has informed much CST research and evaluation in the last decade, including scores related to aspects of engagement, such as immersion, results-worth-effort, and enjoyment. The CSI consists of two parts: i) six eleven-point ordinal item pairs are answered for the creativity-related factors of collaboration, enjoyment, exploration, expressiveness, immersion and results-worth-effort; and ii) fifteen paired comparisons are made across these factors. The total count of factors chosen in the paired comparisons weights the final scores, accounting for which factors are most important in the creative context being assessed. However, the wide range of factors relating to creativity can lead to assessing factors that distract from study goals and inflate the length of user studies.

Another validated questionnaire for assessing engagement in HCI is the User Engagement Questionnaire (UEQ; O'Brien et al., 2018). It is based on the operationalisation of engagement outlined in Section 2.1.2. Unlike the CSI, its factors focus purely on aspects of engagement and not other aspects of creativity support. A short-form version is available, which is helpful as it does not fatigue participants. Metrics are collected for:

- focused attention (points of immersion or concentration with a system),
- aesthetic appeal (the visual appeal of a system),
- perceived usability (how easy a system is to use),
- reward (how fulfilling using a system is),
- and a total engagement score.

These factors are focused on the evaluation of a technology and do not distract from the goal of assessing engagement. However, a critique is that it includes assessments of usability which, whilst related to engagement, can be distinct. For example, systems that are too easy to use can actually support boredom instead of user engagement, cf. the theory of flow (Csíkszentmihályi, 1990).

2.3.2.2 Measuring Reflection

Whilst questionnaires for evaluating engagement have been used often in CST research, questionnaire measures of reflection have not been frequently used. Reflection questionnaires are sparse in HCI, and while reflection questionnaires outside of HCI have been used to examine technology (Levine, 2014; O'Reilly & Milner, 2015; Renner et al., 2014), these are not validated nor widely used. Cox et al. (2025) found that of 173 user studies on CSTs in the last 10 years, only 6 (including research published from this thesis) related to self-reflection.

Education and healthcare researchers developed self-report questionnaires to operationalise reflection, but not for evaluating technology. A systematic review of 700+ papers (Ooi et al., 2021) recommended the Reflection Questionnaire (Kember et al., 2000) and Self-Reflection and Insight Scale (SRIS) (Grant et al., 2002) as most rigorous.

The SRIS informed HCI design considerations for supporting everyday reflection (Mols et al., 2016). However, the SRIS is not technology-focused; instead, it quantifies people's tendency to self-reflect. As recommended by Bentvelzen et al. (2022), SRIS scores are helpful for testing if participants have poor reflective skills that might bias the findings. Metrics are calculated for:

- insight (ability to reach insights),
- engagement in self-reflection (how frequently participants self-reflect),
- need for reflection (whether participants are motivated to reflect),
- and a total SRIS score.

Bentvelzen et al. (2021) developed the Technology-Supported Reflection Index (TSRI) to quantify levels of reflection afforded by personal informatics systems. Their scale provides the measurement closest to the goal of evaluating reflection in CSTs. However, as the TSRI is designed for personal informatics, its questions on long-term personal data do not fit the CST context. Indeed, the TSRI is optimised for interfaces with a functional goal to support people in changing their behaviours given logs of their personal data. This contrasts with creative interfaces where interaction is open-ended and unpredictable (Hewett, 2005). There are no questionnaires to assess reflection during CST interaction.

2.3.3 Data Collection: Interviews and Video-Cued Recall

Questionnaires quantify participants' thoughts and feelings about a technology. However, qualitative approaches are needed to interpret their values and for more nuanced insights (Hewett et al., 2005).

Interviews are frequently used in HCI to give insight into users' thoughts and feelings on their recent interaction with a technology. They can be structured, semi-structured or fully open-ended (Braun & Clarke, 2013). The benefit of more open-ended questions is that they allow researchers to investigate unexpected lines of inquiry relevant to the research question.

However, ensuring that questions do not lead participants towards desired responses is challenging, particularly when participants can misremember aspects of their prior interaction.

The think-aloud method is another qualitative data collection method. In the think-aloud method, participants are asked to describe their thought process whilst performing a task (Preece et al., 2011, pg. 256; Nielsen, 1994, pg. 195). This method gives detailed insight into participants' cognitive processes. It also captures participants' thoughts and feelings at the point of interaction, ensuring greater accuracy than post-hoc approaches, where participants can misremember their experience. However, the think-aloud method can distract users, meaning certain user experience aspects cannot be investigated, such as flow states which occur in distraction-free environments (Csíkszentmihályi, 1990). It also raises concerns about ecological validity – people are unlikely to speak their thoughts aloud when typically interacting with a CST. Thus, the data collected does not represent real-world practice (Candy et al., 2006).

An alternative to the think-aloud method is to conduct a retrospective think-aloud. This technique is also referred to as video-cued recall (Candy et al., 2006; Candy, 2006). In video-cued recall, participants describe *a recording* of their interaction with a CST. Video-cued recall is useful in giving cues to support participants' short-term memory, leading to accounts based on their memory of their cognitive processes (Candy et al., 2006). However, participants can justify their actions post hoc instead of giving accurate insight into their thinking during interaction. Participants are also not trained in identifying moments upon which to reflect; the researcher often has to intervene to select relevant moments for their reflection, thus introducing a bias.

2.3.4 Data Collection: First-person Perspectives

Questionnaires and interviews are used to identify how groups of people interact with technology. Questionnaires reduce a complex construct into measurable categories. The typical analysis of interviews also involves identifying points of interest and grouping these into larger themes (Braun & Clarke, 2006). This can lead to findings which do not capture nuances in an individual's personal responses. Arts and humanities-inspired approaches offer deeper insights into the subjectivity of how individuals interpret their use of technology (Bryan-Kinns & Reed, 2023; Candy, 2011; Candy & Edmonds, 2018).

Subjective viewpoints are captured through first-person accounts: writing by participants or researchers based on their own data collection and experience (Lucero et al., 2019). First-person accounts allow participants to articulate and clarify reflections on their interaction with technology and give insight into their subjective viewpoints on their practice (Höök et al., 2018; Lucero et al., 2019). First-person accounts also give voice to the personal ways that people reflect, generating knowledge that might resonate with readers (Ellis et al., 2011; Fdili Alaoui, 2023). However, this requires that participants have strong enough writing skills to articulate their thinking.

First-person perspectives (Höök et al., 2018; Lucero et al., 2019) have been used in CST contexts, with examples of collaborative pieces (Ben-Tal et al., 2021; Lewis et al., 2023; Sturdee et al., 2021; Sturm et al., 2019), autoethnographies (Bryan-Kinns, Noel-Hirst, & Ford, 2024; Gioti et al., 2022; Lewis, 2023; Lucero, 2018; Noel-Hirst & Bryan-Kinns, 2023; Spiel, 2021; Sturm, 2022) and vignettes (Benjamin et al., 2023; Chang et al., 2023) – all of which demonstrate the value of first-person approaches in giving insights into personal experience in HCI.

2.4 Summary of Literature Critiques

This chapter shows that there is little CST research focused on reflection, and even fewer studies directly assessing reflection as part of a mixed-methods evaluation. There is no method to systematically assess reflection in CST studies and settings. This section closes the chapter by presenting a summary of the critiques of literature.

The first section reviewed interdisciplinary literature to introduce models of the creative process (Amabile, 1983, 1996; Sanders & Stappers, 2008; Sapp, 1992; Wallas, 1926). The critique is that these models provide little insight into the experiential aspects of creative work (Nelson & Rawlings, 2009). Of the many experiential aspects not represented, this thesis focuses on reflection. It is argued that reflection is crucial for creative user experiences, as supported by numerous studies demonstrating people's reflections on their use of creative technology (Candy, 2019; Fdili Alaoui, 2019; Guillaumier, 2016; Lewis et al., 2023; Sturm, 2022).

However, reflection is complex and ill-defined. Many existing definitions (Moon, 2013) are incomplete and lack insight into how reflection develops. Models of the reflection process are also based in fields such as education (Atkins & Murphy, 1993; Boud et al., 1985; Boyd & Fales, 1983; Dewey, 1933; Steinaker & Bell, 1975) and do not relate to creative interaction contexts. Furthermore, outside of the CST field, HCI research that characterises reflection focuses primarily on personal informatics systems (Baumer et al., 2014; Bentvelzen et al., 2022; Dijk et al., 2017). These systems encourage reflection on the functional goal of learning from your own data, whereas CSTs focus on open-ended and ambiguous goals (Hewett, 2005).

Candy (2019) presents the only framework for reflection which focuses on creative contexts. However, Candy's (2019) types of reflection, derived from qualitative interviews, require operationalisation for use in user studies. They cannot be used to quantify and systematically assess reflection in CSTs interaction. In contrast, questionnaire approaches to measuring reflection produce comparable scores. The scores are also gathered directly from users as opposed to being inferences from the researcher. This justifies the need for a quantitative operationalisation of reflection in CST interaction.

The limited number of studies on reflection in the CST domain further supports the need to operationalise reflection in CST interaction. The few CST studies on reflection support mostly reflection back on documentation (Dalsgaard & Halskov, 2012; Kim et al., 2017; Sharmin & Bailey, 2013; Sterman et al., 2023), or present new visualisations of data to encourage reflection on new perspectives (Belakova & Mackay, 2021; Cho et al., 2022; Hoque et al., 2022). These features do not support the tacit (Schön, 1983), in-the-moment (Candy, 2019) reflections that occur during CST interaction; they require post-hoc reflection.

To further justify developing a quantitative operationalisation of reflection for CST interaction, this chapter reviewed the standard methods and techniques for CST evaluation (Hewett et al., 2005). In CST evaluations, self-report questionnaires are typically used, alongside qualitative methods such as interviews or first-person accounts, as part of a mixed-method study design. In particular, questionnaires based on engagement are often used. For example, the influential Creativity Support Index (Cherry & Latulipe, 2014) includes measures for factors of enjoyment and immersion. Existing questionnaires on reflection are developed for non-creative domains such as personal informatics (Bentvelzen et al., 2021), or evaluate people's reflective capacity (Grant et al., 2002) and not their technology use.

A need for a questionnaire for reflection in CST interaction is thus established. This thesis thus contributes a new questionnaire which captures whether more or fewer moments of reflection occurred during interaction with a CST. This enables the systematic evaluation of reflection in different CSTs and study contexts. Its development is documented in the following chapters.

Chapter 3

Design of RiCE Version 1

The previous chapter showed that there is a need for a questionnaire to evaluate reflection in Creativity Support Tool (CST) interaction. This chapter thus documents the initial development of a questionnaire which captures whether more or less moments of reflection occurred during interaction with a CST. The questionnaire in its first version is named the Reflection in Creative Experience Questionnaire version 1 (RiCEv1).

The RiCEv1 development process reported in this chapter is as follows. First, an initial set of questions are devised and then reduced based on scores by people with expertise in creative disciplines (Section 3.2). An Exploratory Factor Analysis (EFA) was then performed to reduce the questions further and group them into factors, based on their scores from 300 recent users of CSTs (Section 3.3). The chapter closes with a discussion of RiCEv1's factors (Section 3.4) in relation to the current literature.

3.1 Method Overview

Methods from existing questionnaire literature were consulted to develop RiCEv1 (see Section 2.3.2). First, a set of items was generated based on background reading and reduced via reviews by the thesis's author, their supervisor and experts in creativity. Second, an EFA reduced the items into factors based on 300 people's recent experiences with a CST.

The method is similar to the Technology Supported Reflection Index (TSRI; Bentvelzen et al., 2021). However, the analysis is applied to creative tasks inspired by the studies conducted for the Creativity Support Index (CSI; Cherry and Latulipe, 2014). This is because there is no consensus on which aspects of reflection are most valuable in creative contexts. Thus, the items are developed and categorised into factors statistically (as in the TSRI) instead of matching items to factors beforehand (as in the CSI). The goal was also to examine RiCEv1 in contexts similar to CSI, as it is frequently used to evaluate CSTs.

The study was approved by the Queen Mary University of London's ethics committee. Participants were fully briefed, reimbursed for their time and gave written consent. See the appendix for consent forms, questionnaires, data collected and code written in the R programming language[1] for its analysis.

3.2 Item Development

The first stage was to develop a list of items that capture moments of reflection in a recent CST interaction. A quantitative approach was followed where people with expertise in creative disciplines (defined in Section 3.2.2) scored items independently. This approach is relatively quick as multiple people would not meet to debate nuances, as with a qualitative approach, respecting their limited time as full-time working creative professionals.

The following subsections detail how the items were developed. Section 3.2.1 details preliminary work used to develop RiCEv1's initial items. These items were then assessed by creative professionals as described in Section 3.2.2.

3.2.1 Preliminary Work

The author first searched through items from existing measures used to evaluate CSTs (Cherry & Latulipe, 2014; Jackson et al., 2008; O'Brien et al., 2018) and measure people's reflection (Bentvelzen et al., 2021; Grant et al., 2002; Kember et al., 2000; Levine, 2014; O'Reilly & Milner, 2015; Renner et al., 2014). An item is defined as a statement to be rated by people against a row of ordinal points.

[1] https://www.r-project.org

Candidate items were sorted into the aspects of reflection listed in Section 2.1.1.1, acting as a guide for whether items could indicate moments of reflection. 62 items were rephrased to relate more directly to creativity and reflection. 49 novel items were written based on the PhD's literature review, including recent CST studies discussing reflection (see Section 2.2.1). 115 items were created in total. Three examples are shown in Table 3.1. The full list is in the appendix.

Table 3.1: Three examples of items initially generated for the RiCEv1 design. See the full list in the appendix.

Item	Aspect	Comment
The system worked in ways which were often puzzling (Grant et al., 2002).	Breakdown	Modified from SRIS to be system oriented.
I identified connections between contrasting ideas and explored this in my creation (Yurman, 2021).	Comparison	Novel item.
I was able to easily explore other people's ideas. (Bentvelzen et al., 2021)	Openness	Modified from TSRI to not focus on data.

To reduce the item set, the thesis's author and their supervisor independently scored each item as "Disagree" (1), "Neutral" (2), or "Agree" (3) against the criterion:

> "The item appropriately contributes towards assessing if a moment of reflection occurred during a person's creative user experience."

As some items fit multiple aspects of reflection, the items were shuffled and presented without categorisation; the statistical analysis in Section 3.3 drives item groupings. A Cronbach's (1951) alpha, which is suitable for assessing agreement between raters when using ordinal data, of .76 was calculated. This shows acceptable agreement between raters following general guidelines (Schrepp, 2020).

The author and supervisor discussed items where their scoring contrasted. The set was then reduced by removing 60 items where at least 1 rater scored "Disagree", excluding 4 items where wording was tweaked. This resulted in 59 items being shuffled and scored again by the raters independently, against a revised criterion statement of:

> "The item indicates that a moment of reflection occurred whilst a person was undertaking a creative activity with technology."

A Cronbach's (1951) alpha of .71 was calculated; there is acceptable agreement between raters (Schrepp, 2020). Of the 59 items, 37 where there was full agreement between raters were assessed by 10 creative professionals, as described below.

3.2.2 The Creative Professionals' Review

Ten creative professionals were recruited through the author's professional network of PhD students and academics. Creative professionals are defined here as people with knowledge of the creative process, where some experience with creativity-related HCI or knowledge on reflection is desirable. The sample size of 10 was chosen because Boateng et al. (2018) outline that typically 5 to 7 expert evaluators are used to develop questionnaires, rounding upwards for simplicity. Effort was made to represent many creative disciplines to identify items that are useful to many CST researchers. Table 3.2 shows the creative professionals' genders (4 Male, 6 Female), ages (Mean = 28.2, Med = 28, SD = 4.29), countries and summaries of their self-written biographies. They are referred to as P1 to P10 below.

3.2.2.1 Procedure

The creative professionals were sent a spreadsheet with the 37 items devised in Section 3.2.1 and instructions for scoring. They were asked to score items "Disagree" (1), "Neutral" (2), or "Agree" (3) against the criterion refined in the preliminary work:

> "The item indicates that a moment of reflection occurred whilst a person was undertaking a creative activity with technology."

A notes column was also provided, where the creative professionals were encouraged to give further feedback. Items were shuffled for each creative professional. Each creative professional was reimbursed with a £20 Amazon voucher for their time. The procedure lasted between 30 and 45 minutes.

Table 3.2: Creative professionals' backgrounds for the RiCEv1 design. Biographies are summarised from verbatim biographies found in the appendix. All participants were instructed to write their biographies to only include information that they consent to be published, as approved by the Queen Mary University of London's ethics committee.

ID	Age	Gender	Country	Biography Summarised
P1	29	Female	China	Final year PhD; Musical Interaction; Digital Musical Instrument design; MArch Urban Design; BEng School of Architecture; Teaching experience related to creativity, design and applying technology in these fields.
P2	23	Female	Italy	End of first year PhD in AI and Music; Attended conservatoire for piano performance and composition; A-Level Music Technology; Creative Music Technology degree.
P3	34	Female	England	Second year PhD in Computational Creativity; Examining text-to-image generative AI and Twitter bots; MSc Computer Science; BA(Hons) Fine Art; self-employed (tattoo) artist for several years; ProCreate; Photoshop; Produced paintings for exhibitions.
P4	27	Female	Germany	First year PhD in the Art and Design faculty; Research Assistant in the Computer Science faculty; background in Industrial and Interaction Design; Mentor for first year university students, guiding reflective practices.
P5	33	Female	England	Fourth year PhD; Exploring mindfulness in Interaction Design with AI and Audio.
P6	34	Male	Chile	Third year PhD in Media and Arts Technology; Researching error and music improvisation; experience in web development; Multi-instrumentalist: piano, voice, guitar, venezuelan cuatro; performer and composer.
P7	25	Male	England	Associate Lecturer in Music Technology; BSc(Hons) Music Technology; MSc Creative Technology; Composer of punk and hard rock/metal through to alt-jazz; experience with p5.js and openFrameworks, Unity, Unreal, MaxMSP and Ableton.
P8	29	Female	USA	Fifth year PhD in HCI; Investigating Human-AI Co-Creativity, Ethical AI and Interaction Design; BSc Computer Science and Engineering; Teaching experience in HCI and rapid prototyping.
P9	24	Male	England	Award winning filmmaker; Short films, animation and live action, telling stories on South Asian experiences; Storyboarder; Celtx; Fade-In; Adobe CC Suite (After Effects, Premiere Pro); Davinci Resolve Studio; Final Cut; Clip Studio and TV Paint.
P10	24	Male	Norway	Assistant Film and TV Colourist in a post-production house; VFX turnovers; Grade-matching; Experience working on music videos, short films and TV Series; Baselight; DaVinci Resolve; Premiere Pro.

3.2.2.2 Analysis Method

For each item, “Disagree”, “Neutral” and “Agree” responses were counted. The items were then listed and ordered by the number of “Agree” responses to compare and contrast the highest and lowest scoring items. Furthermore, the scoring was interpreted in the context of the creative professionals’ comments. Items for the next phase were retained where more than 7 out of 10 creative professionals selected “Agree”. Inter-rater reliability is calculated and interpreted using Cronbach’s (1951) alpha as in Section 3.2.1.

3.2.2.3 Results

For the creative professionals’ scoring of all items, Cronbach’s (1951) alpha equals .72. This shows acceptable agreement between raters (Schrepp, 2020). Table 3.3 lists the highest and lowest scoring items sorted by the number of “Agree” responses. The horizontal line indicates where items are omitted for brevity.

Some items with high “Agree” scores relate to iterating (Q23, Q7), self-assessing and selecting actions (Q14, Q11, Q13, Q29). P7 noted that “you can reflect on each interaction to understand why each may not have worked”. P10 noted they are “constantly learning and refining techniques”. This shows that a cyclical process of improvement is important to reflection in creative work.

Items concerning how others perceive your creative work (Q24, Q27, Q33) scored low. P1 said that “if the creative activity is about self-expression”, worrying about others’ perceptions might not indicate reflection. P10 did not “mind what others [thought]”. This shows that moments of reflection in creative activities are personal to creators. Indeed, some high-scoring items relate to personal improvement (Q1, Q19, Q21). Furthermore, creative professionals scored low on items related to their beliefs being challenged (Q9, Q26). P3 wrote “being challenged != reflecting”, whereas P4 noted that such items “better suit reflexivity” than reflection.

Table 3.3: Count of creative professionals scoring "Agree", "Netural" or "Disagree" for select items, sorted by the number of "Agree" scores. Items where 7 out of 10 or more creative professionals rated "Agree" were taken forward to the scale development phase. (R) denotes that the item's answer given by a participant in a user study should be reversed for analysis.

Q	Item	Total Count		
		"Agree"	"Netural"	"Disagree"
Q23	I often generated, tested and revised ideas.	10	0	0
Q25	Whilst creating, I thought back on some of my past experiences.	10	0	0
Q30	I often reflected on my actions to see whether I could have improved on what I did.	10	0	0
Q7	I found myself iteratively refining and assessing my creative process.	9	1	0
Q14	I pondered over the meaning of what I was doing in relation to my personal experiences.	9	1	0
Q1	I constructively self-assessed my own actions.	9	0	1
Q12	Whilst being creative, it was very interesting to examine different aspect of my creation.	9	0	1
Q5	I sometimes felt doubtful whilst creating my project.	8	2	0
Q11	I made comparisons within the system to consider alternative ways of doing things.	8	2	0
Q13	Whilst being creative, I liked to think about my actions to find alternative ways of doing them.	8	2	0
Q22	I explored my past experiences as a way of understanding new ideas.	8	2	0
Q29	I considered different ways of doing things.	8	1	1
Q2	I considered how my outputs from the system might be interpreted differently in the future	8	1	1
Q35	I often re-examined things I'd already learnt.	7	3	0
Q19	I learned many new things about myself during the experience.	7	2	1
Q21	I often reappraised my experiences with the system so I could learn from them.	7	2	1
Q24	I was not worried about what others may have been thinking about me (R).	3	5	2
Q32	The results of my actions often violated my expectations.	3	3	4
Q27	I didn't really think about how others would perceive my creative process and final product. (R)	3	1	6
Q33	I was not concerned with how others might evaluate my performance (R).	2	5	3
Q26	The system challenged some of my firmly held beliefs.	2	4	4
Q9	Some of my firmly held beliefs were challenged.	1	5	4

3.3 Scale Development

In the previous phase, a set of 115 items was reduced to 16 items that indicate a moment of reflection during a CST interaction. The following describes an online survey subjecting these 16 items to an EFA to group them into factors.

3.3.1 Participants

Participants were recruited using Prolific[2], an online survey platform. Prolific was used because it is academic-focused and its users show interest in creativity-related work (Oppenlaender et al., 2020). Prolific's pre-screening features were applied to distribute the survey to people worldwide who reported being fluent in English, had a Prolific approval rating above 98%, and used a device with a screen at least weekly. Participants were also required to have used a CST within the last 2 weeks. In the study, the summary of CSTs from Cherry and Latulipe (2014, pg. 3) was given as examples to participants. Participants could also self-report their own CST to consider whilst completing the survey.

Recruitment was continuous until 300 participants were reached after data cleaning (see Section 3.3.3). This was because Boateng et al. (2018) outlined that multiple authors describe n = 300 as "good" for factor analysis. 320 participants were recruited in total, rejecting 20.

Participant genders collected in response to the open question "What is your gender?" were: 56.3% Male, 41.3% Female, 1.6% Non-Binary, 0.3% Trans Man and 0.3% None (which is taken to mean 'prefer not to say'). Mean age was 29.1 (Med = 26, SD = 9.19). Figure 3.1 shows the participants' countries of origin. Most participants are from Portugal (21.3%), South Africa (18.0%), the UK (12.3%) and Poland (12.3%). Participants were reimbursed an average of £9.52/hr. The survey took a mean of 10m 55s to complete (Med = 09m 41s, SD = 5m 24s).

[2] https://www.prolific.co

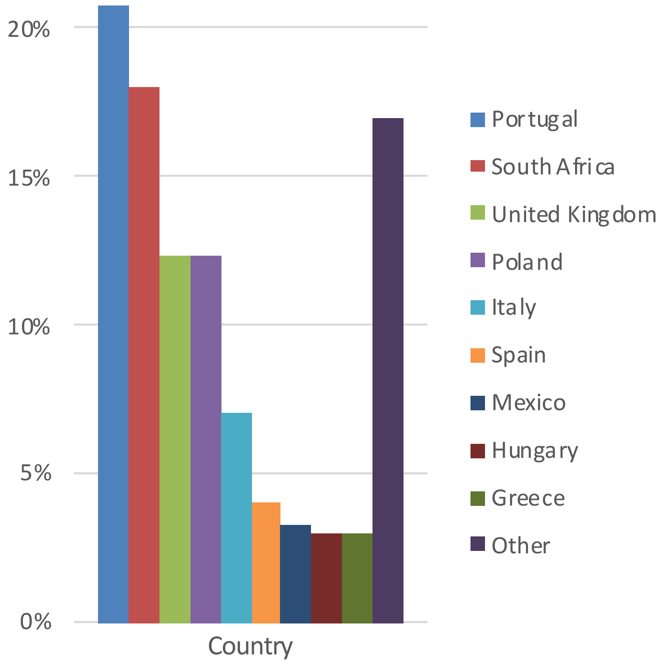


Figure 3.1: Participants' countries of origin in the RiCEv1 design study.

3.3.2 Procedure and Data Collection

The measures collected in the online survey were as follows:

1. **Pre-test Questionnaire:** Demographic information was collected as reported above. Participants also completed the Self-Reflection and Insight Scale (SRIS) scale (Grant et al., 2002) to evaluate if the sample has a bias in its natural tendency to self-reflect (see Section 2.3.2.2).

2. **Creative Technology:** Participants were asked to select "a creative technology which [they] have used in the last 2 weeks". To recap, a drop-down list was provided in the survey based on the summary table of CSTs in Cherry and Latulipe (2014, pg. 3) but participants could also respond with a free-text description of their own technology. They were then asked to:

   > "Briefly describe below the creative technology that you have chosen above, how you use it, how it supports your creativity, and how it supports creativity in general."

This description was used to clean the data and check participants' understanding of their chosen technology (see Section 3.3.3).

3. **RiCEv1:** Participants were shown the 16 items identified in the creative professionals' review. They were instructed to rate them "considering their recent experience with their selected creative technology". Each item was placed alongside an 11-point scale with the anchors "Highly Disagree" (0) and "Highly Agree" (10) on either end. These anchors directly mirror the CSI; as it is popular for CST research, RiCEv1 will often be used alongside it. 11 points were chosen as multipoint items are often easier to use (Lewis & Erdinç, 2017) and more points supports test-retest reliability (Preston & Colman, 2000).

Participants could also add further comment in an open-ended text box.

3.3.3 Data Cleaning

The data cleaning procedure followed the advice in Müller et al. (2014). First, participants' understandings of their chosen technology were checked via an open-ended question (see Section 3.3.2); 6 participants were removed who said they had not used a creative technology or did not describe their chosen technology in sufficient detail. Second, duplicate responses were checked; no responses were identical. Third, a histogram of the survey completion times was examined to identify outliers, removing 6 participants who spent longer than 30 minutes. Fourth, 8 "flat-liners" (Müller et al., 2014) were rejected who had selected the same option for all items in at least one question block. Respondents were required to complete each question before submission; there was no missing data. This led to the 300 participants (20 out of 320 completed surveys were removed).

3.3.4 Data Analysis Method

The choice of creative technology and SRIS scores is reported as descriptive statistics. For the EFA, Taherdoost, Sahibuddin and Jalaliyoon's (2014) advice is followed. Firstly, the sample adequacy is assessed by determining whether the Kaiser-Meyer-Olkin (KMO) value is $\geq .7$ (Kaiser, 1970). Next, Bartlett's (1950) test of sphericity is tested for significance ($p < .05$) to indicate that correlations between items are large enough for factor analysis.

If these tests are passed, an EFA with the minimum residual method (Lloret-Segura et al., 2014) and oblique rotation is used because, as with the CSI, RiCEv1's items are correlated (Cherry & Latulipe, 2014). Next, the number of factors is identified where Eigenvalues are > 1.0; this shows that each factor has a higher variance compared to a single item (Taherdoost et al., 2014). Then, for each valid factor, Kaiser's (1960) rule is followed to select items uniquely correlating with (or loading onto) said factor $\geq .4$.

Furthermore, Cronbach's (1951) alpha is calculated to assess inter-item reliability (if items in each factor measure similar constructs), following the guideline that alpha values $\geq .7$ are acceptable. However, the interpretation of the alpha value is lenient as scales with few items per construct will naturally yield lower alphas (Schrepp, 2020), and RiCEv1 aims to be lightweight.

3.3.5 Results

Table 3.4 shows the CSTs participants chose when answering the questionnaire. This included software for writing, presentations, photo editing and programming.

Table 3.4: Counts of participants' choice of creative technology in the RiCEv1 design.

No. Participants	Creative Technology
20+	MS Word (43); Photoshop (42); Google Docs (29); MS Powerpoint (24)
10+	Visual Studio (15); Adobe Lightroom (15); Blender (13); Adobe Premier Pro (11); AutoCAD (10)
5+	WordPress (8); Google Slides (8); MatLab (7); Illustrator (6); iMovie (6); Paper and Pen (5)
3+	Unity (4); Post-It Notes (3); R Studio (3); Cubase (3)
2	Tableau; Whiteboards; WolframAlpha; Scratch; Final Cut Pro; Adobe After Effects; GarageBand; Prezi; Mendeley; MS Publisher; Cinema 4D; Canva
1	XCode; InkScape; CorelDraw; Logic Pro X; Wikis; MediaWiki; DreamWeaver; Celtx; Obsidian; Clip Studio Paint; Figma; Ableton Live; Arduino; Bear; Kdenlive; Power BI; GIMP; FL Studio; Procreate; TV Paint

The SRIS scores are shown in Figure 3.2. This shows that participants are motivated to engage in reflection but do not always understand their insights.

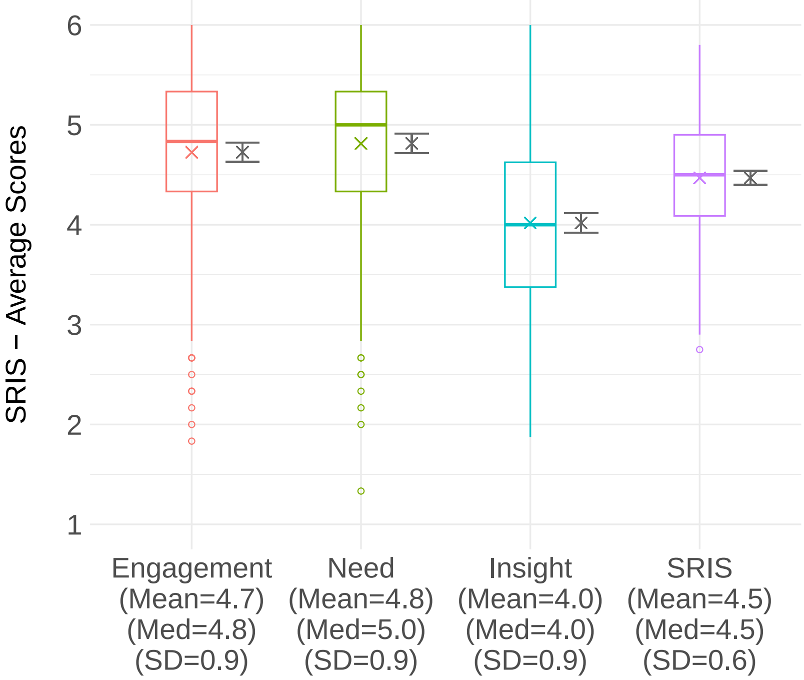


Figure 3.2: SRIS scores in the RiCEv1 design study.

The sampling adequacy was acceptable (KMO $=$ .90) and Bartlett's test of sphericity was significant ($\chi^2(120) = 2110.18$, p $<$ .001); the criteria for conducting a factor analysis is met. Table 3.5 shows the loadings for 4 factors with Eigenvalues $>$ 1 explaining 54% of variance. It also shows the items loading onto a single factor. Factors 1 through 4 explain 17%, 13%, 11% and 13% of variance respectively. As only 2 items loaded onto factor 3 $\geq$ .4, the top 2 highest loading items from each factor are selected.

The decision to select 4 factors with 2 items each is also because: this follows the CSI's (Cherry & Latulipe, 2014) format, RiCEv1 aims to be as short as possible to minimise participants' fatigue, inspecting the EFA with only 3 factors to items per factor identified groupings that were harder to interpret (Worthington & Whittaker, 2006), and 5 factors did not achieve the necessary Eigenvalues.

Table 3.5: **Loadings for the items in the RiCEv1 design. Values $>$ 0.4 are in bold.**

Question	Single Factor	Factor 1	Factor 2	Factor 3	Factor 4
Eigenvalue	6.00	2.68	2.05	1.69	2.02
Q11) I made comparisons within the system to consider alternative ways of doing things.	**0.54**	0.00	0.07	0.02	**0.66**
Q23) I often generated, tested and revised ideas.	**0.51**	0.05	-0.16	0.20	**0.57**
Q30) I often reflected on my actions to see whether I could have improved on what I did.	**0.68**	0.38	0.00	0.00	**0.49**
Q12) Whilst being creative, it was very interesting to examine different aspect of my creation.	**0.76**	0.33	0.16	0.12	0.37
Q29) I considered different ways of doing things.	**0.67**	**0.50**	-0.05	0.09	0.28
Q35) I often re-examined things I'd already learnt.	**0.66**	**0.65**	0.08	0.09	-0.02
Q13) Whilst being creative, I liked to think about my actions to find alternative ways of doing them.	**0.76**	**0.88**	0.02	0.03	-0.02
Q7) I found myself iteratively refining and assessing my creative process.	**0.72**	**0.42**	0.12	0.10	0.26
Q1) I constructively self-assessed my own actions.	**0.72**	0.36	0.23	0.08	0.25
Q22) I explored my past experiences as a way of understanding new ideas.	**0.68**	0.12	-0.01	**0.81**	-0.05
Q25) Whilst creating, I thought back on some of my past experiences.	**0.57**	-0.10	0.08	**0.73**	0.06
Q5) I sometimes felt doubtful whilst creating my project.	0.30	0.06	0.20	-0.13	0.29
Q2) I considered how my outputs from the system might be interpreted differently in the future.	**0.47**	-0.06	**0.54**	-0.02	0.26
Q14) I pondered over the meaning of what I was doing in relation to my personal experiences.	**0.49**	-0.14	**0.61**	0.20	0.09
Q19) I learned many new things about myself during the experience.	**0.46**	0.08	**0.79**	0.00	-0.13
Q21) I often reappraised my experiences with the system so I could learn from them.	**0.62**	0.20	**0.57**	0.08	0.03

Given this, RiCEv1 is presented in Table 3.7. The factor names were based on discussions between the thesis's author and their supervisor. Table 3.7 also includes Cronbach's (1951) alpha values showing acceptable to moderate inter-item reliability between all factors. Moderate factors were retained as alpha was calculated for only 2 items, making a low value probable (Schrepp, 2020). These items also scored highly in the creative professionals' review (see Table 3.3) and are interpretable in a conceptually meaningful way (Worthington & Whittaker, 2006).

3.4 Discussion

This chapter detailed the initial design of a lightweight questionnaire to evaluate reflection in CST interaction. Table 3.7 shows the first iteration of RiCE, RiCEv1. Its design is based on prior literature, a review of items by creative professionals and an EFA. A summary of the main findings of this chapter is shown in Table 3.6. In this section, RiCEv1's factors are discussed below (see Section 3.4.1) in relation to current literature (see Chapter 2). Limitations are then outlined in Section 3.4.2. RiCEv1's factors are referred to using the dimension identifiers in Table 3.7.

Table 3.6: Summary of Chapter 3's main findings.

Main Finding	Location
RiCEv1 characterises reflection in CST interaction as four factors: reflection-on-self, reflection-through-experimentation, reflection-on-process, and reflection-on-past. Each is quantified by two questionnaire items per factor and understandable in a conceptually meaningful way.	§3.3.5
RiCEv1 characterises reflection in CST interaction as a cyclical process based on users' in-the-moment judgments and intuitions. This is based on a review of questionnaire items by creative professionals.	§3.2.2.3
RiCEv1 characterises reflection in CST interaction as including moments where users consider their personal self-expression. A review of questionnaire items by creative professionals found that they considered reflection on others' judgments less important than self-reflection in CST interaction.	§3.2.2.3

Table 3.7: Items and instructions for administering and scoring RiCEv1. Cronbach's (1951) alpha values are also reported.

RiCEv1: VERSION 1

::: INSTRUCTIONS FOR ADMINISTERING :::

When administering RiCEv1, each item should be placed along an 11-point scale from "Highly Disagree" (left) to "Highly Agree" (right). Values for each item are zero indexed, i.e., integers from 0 to 10. Please follow the question wording exactly, replacing only the name of your system where indicated. Dimension identifiers (e.g. Cp1), descriptions, and headings should not be visible to participants. Item order should be randomised.

Considering your recent experience of [SYSTEM], please indicate the extent to which you agree with the following statements:

Factor 1 (RiCEv1-Cp): Reflection on Current Process ($\alpha = 0.79$)

Cp1 (Q13): Whilst being creative, I liked to think about my actions to find alternative ways of doing them.

Cp2 (Q35): I often re-examined things I'd already learnt.

Factor 2 (RiCEv1-Se): Reflection on Self (α=0.68)

Se1 (Q19): I learned many new things about myself during the experience.

Se2 (Q14): I pondered over the meaning of what I was doing in relation to my personal experience

Factor 3 (RiCEv1-Pa): Reflection on Past Experiences (α=0.77)

Pa1 (Q22): I explored my past experiences as a way of understanding new ideas.

Pa2 (Q25): Whilst creating, I thought back on some of my past experiences.

Factor 4 (RiCEv1-Ex): Reflection through Experimentation (α=0.65)

Ex1 (Q11): I made comparisons within the system to consider alternative ways of doing things.

Ex2 (Q23): I often generated, tested and revised ideas.

All items $\alpha = 0.79$.

::: INSTRUCTIONS FOR SCORING:::

Following the design of related questionnaires (Cherry & Latulipe, 2014; Grant et al., 2002; O'Brien et al., 2018), the total RiCEv1 score (out of 10) is calculated as (Cp1+Cp2+Se1+Se2+Pa1+Pa2+Ex1+Ex2) $\div$ 8. Each of the 4 factors are calculated as the sum of its items divided by 2. For example, Reflection on Current Process is (Cp1+Cp2) $\div$ 2.

3.4.1 RiCEv1's Factors

The creative professionals' review shows that moments where people iterate and continually assess their ideas characterise reflection in CST interaction. This corroborates Dewey (1933) and Baumer's (2015) characterisations of inquiry processes in reflection, and CST researchers (Cho et al., 2022; Guillaumier, 2016; Yurman, 2021) who characterise how people refine their creative work. The inclusion of the RiCEv1-Cp factor also shows that people adapt their creative processes upon reflection. This corroborates the transformation stages in models of reflection (Atkins & Murphy, 1993; Boud et al., 1985; Boyd & Fales, 1983).

The creative professionals' review also supports Norman (1993) and Bentvelzen et al.'s (2022) characterisations that people make comparisons when reflecting. The creative professionals rated items highly on making comparisons to past experiences (see Table 3.3, Q25, Q14, and Q22). The EFA shows that people make comparisons between their personal experiences (RiCEv1-Se), past experiences (RiCEv1-Pa), and as part of (RiCEv1-Ex) and looking back on their current process (RiCEv1-Cp). The distinction between RiCEv1-Ex and RiCEv1-Cp is similar to Schön's (1983) reflection-in-action (making comparisons between ideas during the creative process cf. RiCEv1-Ex) and reflection-on-action (looking back on one's creative process more broadly, cf. RiCEv1-Cp). This also shares characteristics with Candy's (2019) distinction of reflection-in-the-making-moment. The difference, however, is that reflections-on-past occurred during the creative user experience and not as a distinct post-hoc stage.

The creative professionals gave little mention of reflections on breakdowns (Baumer, 2015) or reflection-on-surprise (Candy, 2019). Instead, the creative professionals' scoring and RiCEv1's self-reflection factor (RiCEv1-Se) emphasise "self-expression" (P1); the contemplation of others' perceptions occurs infrequently. This demonstrates that Fleck and Fitzpatrick's (2010) characterisation of reflection (considering broader impacts), or selecting ideas corroborating with a consensus (Dewey, 1933; Fleck & Fitzpatrick, 2010), is less important for the participants in this study than their intuition. This finding also contrasts with the TSRI (Bentvelzen et al., 2022), which found that comparing one's data with another's prompts reflection. However, the study design influences the unimportance of others' perspectives because the participants scored the items alone. They only considered their own creative practice, not creative work for clients (Candy, 2019) or collaborative contexts.

3.4.2 Limitations

In the scale development phase, seven or more creative professionals fully agreed that the items in RiCEv1 capture reflection. This demonstrates that RiCEv1 reliably includes measures understood as relating to reflection. Notably, the EFA identified factors that, considered alongside the discussion of RiCEv1's factors above and related work, are interpretable in a conceptually meaningful way (Worthington & Whittaker, 2006). However, the creative professionals' descriptions of their background show that RiCEv1's factors are biased towards music – six out of ten creative professionals worked with audio in some form (see Table 3.2).

From Slovak, Frauenberger and Fitzpatrick's (2017) perspective, a critique is that RiCEv1's factors are too practitioner-centred. They do not directly indicate whether aspects of a technology-supported environment encouraged reflection. However, RiCEv1 instead focuses on the user experience. It can be applied to identify which aspects of a technology support reflection by administering the questionnaire in different conditions and testing different interface designs. However, the questions themselves are not related to interface elements.

3.5 Conclusion

The previous chapter showed that there is a need for a questionnaire to evaluate reflection in CST interaction. This chapter thus documents the initial development of a questionnaire which captures whether more or less moments of reflection occurred during interaction with a CST. The questionnaire is named the Reflection in Creative Experience Questionnaire (RiCEv1). RiCEv1 consists of four factors: reflection-on-process, reflection-on-self, reflection-through-experimentation and reflection-on-past. These factors are conceptually meaningful and can be applied to systematically evaluate CSTs. The next chapter evaluates RiCEv1 to assess its reliability in a CST user study context.

Chapter 4

Evaluation of RiCE Version 1

The previous chapter detailed the development of the Reflection in Creative Experience Questionnaire Version 1 (RiCEv1). RiCEv1 is found in Table 3.7. This chapter presents a user study to evaluate RiCEv1's reliability in two creative Human-Computer Interaction (HCI) contexts related to creative writing and music making. These areas are chosen to examine RiCEv1 in contexts similar to its intended use, before it is applied to the thesis case study domain from Chapter 6 onwards.

The method (Section 4.1) and findings (Section 4.2) are presented below, followed by a discussion of the findings and their limitations (Section 4.3).

4.1 Method

The method used to evaluate RiCEv1 follows standard practice for HCI questionnaire development (Boateng et al., 2018). In summary, two novel Creativity Support Tools (CSTs) are compared across two different time points. As in Chapter 3, the method follows the Technology Supported Reflection Index (TSRI; Bentvelzen et al., 2021), but analysis is applied to creative tasks inspired by the studies conducted to assess the Creativity Support Index (CSI; Cherry and Latulipe, 2014).

The study was approved by the Queen Mary University of London's ethics committee. Participants were fully briefed and gave consent. Data and analysis materials are found in the appendix.

4.1.1 Participants

58 participants were recruited through Prolific. 54 participants returned to repeat the study procedure 1 week later. Prolific was used due to its suitability in creativity-related academic research (Oppenlaender et al., 2020). The sample size was based on an a priori calculation in the software G*Power for the Wilcoxon signed-rank test because the data collected is ordinal and within subjects (effect size = .5, alpha = .05, power = .95, two-tailed), plus 1 more participant to balance groups.

Participants were screened for those who reported being fluent in English and with an approval rating above 98%. Unlike in Section 3.3.2, participants were not required to have previous experience with a CST. As participants are provided with novel interfaces, they all have the same level of familiarity. Descriptive statistics for participants' age, gender, compensation and time spent are in Table 4.1. Figure 4.1 shows the percentage of participants from each country for both the initial answering of the study and its repetition 1 week later.

4.1.2 User Interfaces

RiCEv1 aspires to be used in many creative domains. Therefore, two interfaces were developed for testing. The interfaces contain aspects of writing, music and drawing to represent typical CST activities (Cherry & Latulipe, 2014; Frich et al., 2019). The interfaces are simplistic, including the minimal number of features required for people to have a short, creative user experience. This focuses the study on the creative task, not other factors such as learning effects (Bryan-Kinns & Reed, 2023).

Existing tools were not used as they might require lengthier learning processes. For consistency, all participants were required to have no prior experience with the interfaces. Furthermore, many CST studies focus on evaluating novel high-fidelity prototypes instead of interfaces with a longstanding release (Frich et al., 2019). This makes novel CSTs an appropriate subject for this investigation. The interfaces were developed with the p5js JavaScript library (McCarthy et al., 2015) and embedded into the questionnaire alongside descriptions of how to use them, requiring no installation.

Table 4.1: Participants' descriptive statistics for the RiCEv1 evaluation.

Test (n=58)			
Gender		**Male:** 43.1%	**Female:** 56.9%
Compensation			£9.89/hr
	Mean	**Med**	**SD**
Age	27.57	25	8.92
Time Spent	18m 49s	15m 6s	9m 33s
Re-test (n=54)			
Gender		**Male:** 44.1%	**Female:** 55.9%
Compensation			£10.80/hr
	Mean	**Med**	**SD**
Age	27.89	25.5	9.13
Time Spent	16m 55s	15m 5s	7m 35s

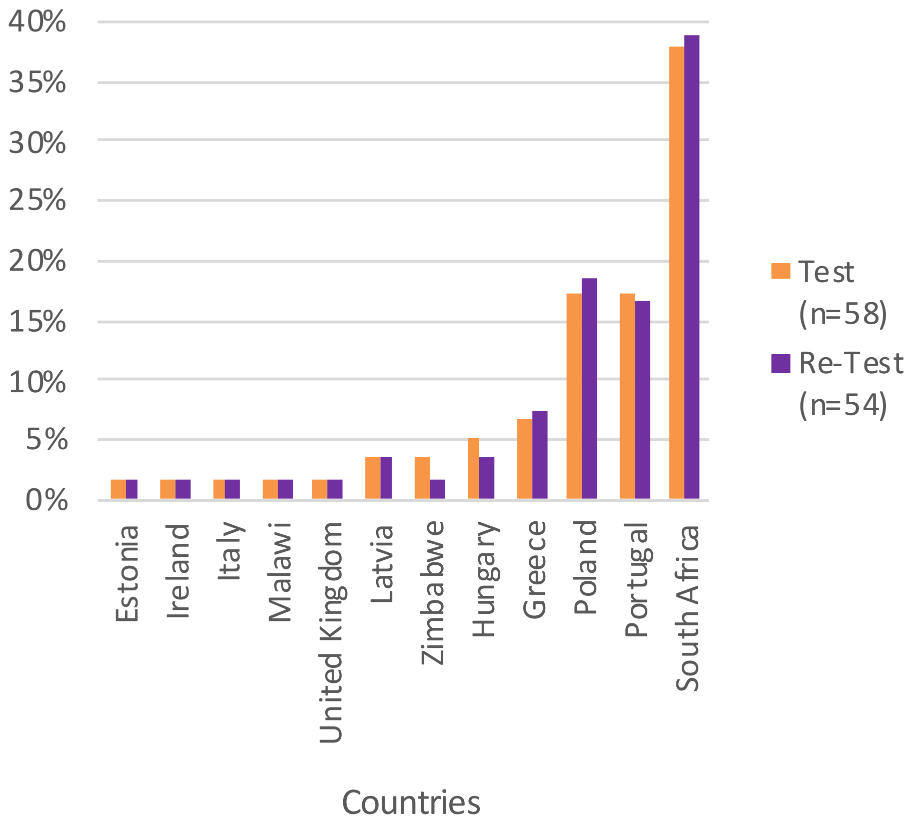
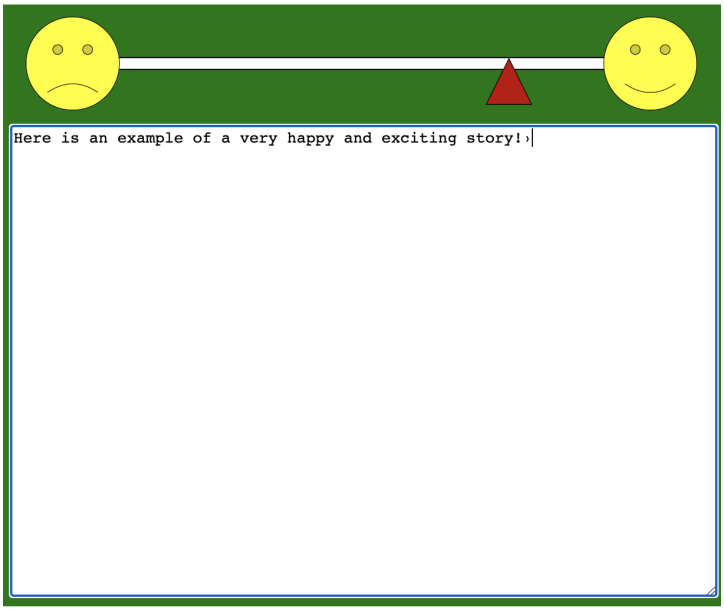


Figure 4.1: Participants' countries of origin in the RiCEv1 evaluation.

4.1.2.1 Story-Sentiment-Visualiser

In *story-sentiment-visualiser*[1], shown in Figure 4.2, people are given real-time feedback whilst writing. As text is typed into the interface, each word is allocated a valence score (positive or negative) based on the AFINN-111 dataset (Nielsen, 2011). This score is visualised by moving the arrow on the smiley scale at the top of the interface and changing the background colour from red (for negative values) through to green (for positive values).



Here is an example of a very happy and exciting story!›

Figure 4.2: Interface for story-sentiment-visualiser. Participants were tasked with writing a positive story for two minutes.

Story-sentiment-visualiser aims to be an interface involving aspects of a creative activity (writing) and aspects of a reflective CST (see Section 2.2.1). Its design is inspired by principles on how to design technology for reflection. For example, the visual feedback provides more information than people usually see while writing, cf. Fleck and Fitzpatrick (2010) and the CSTs that offer new perspectives in Section 2.2.1. Participants using this interface were tasked with writing a positive story so that their intent is visualised, cf. Kreminski and Mateas (2021), for two minutes.

[1] https://codetta.codes/story-sentiment-visualiser/

Creative writing was examined because the CSI (Cherry & Latulipe, 2014) was also tested in this domain. Creative writing is also an area where reflection has been investigated (Belakova & Mackay, 2021; Carrera & Lee, 2022; Hoque et al., 2022; Kreminski & Mateas, 2021). The task requires little prior knowledge, making it suitable for novices to achieve in a short time frame.

4.1.2.2 Sound-sketcher

Sound-sketcher[2], shown in Figure 4.3, allows people to draw points which are sonified into a melody. X-coordinates equal time and y-coordinates equal pitch. People can play and stop the sonification using the play button in the top left corner. Their composition is not played in real-time but only when the play button is clicked. Users can also switch between a pen and eraser tool, the latter allowing them to remove points.

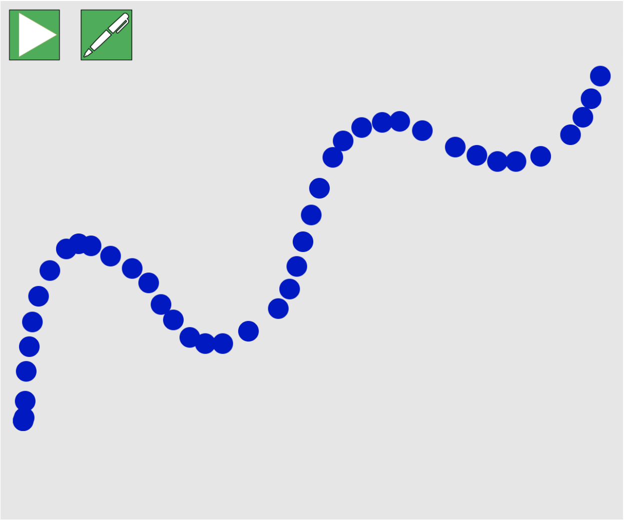


Figure 4.3: Interface for sound-sketcher. Participants were tasked with writing a music composition for two minutes.

Inspiration was drawn from tools that support novices' music making by converting drawings into sound (Dannemann & Barthet, 2021; Farbood et al., 2004; Löbbers et al., 2021; Thiebaut et al., 2008). Sound-sketcher aimed to examine a more open-ended and ambiguous (Gaver et al., 2003) CST design than with story-sentiment-visualiser, and to investigate elements of

[2] https://codetta.codes/sound-sketcher/

music CSTs as per the thesis's later case study. It was decided that this style of tool would examine the intersection of both music and sketching domains, whilst acknowledging that this is an oversimplification; music and sketching are distinct and broad areas of which sound-sketcher only captures some characteristics. The tool also allows people to create music relatively quickly. Participants were tasked with composing a piece of music for two minutes.

4.1.3 Procedure and Data Collection

The following questionnaire measures were used and presented in the order listed below:

1. **Pre-test Questionnaire:** Demographics were collected as reported in Section 4.1.1. Participants also completed the SRIS (Grant et al., 2002) to assess people's capacity for self-reflection; a total average score is calculated.

2. **Tasks:** Participants first use one of the interfaces to complete its associated task. Later, the participants use the other interface to complete its associated task. The order is randomised but balanced (50% started with sound-sketcher, 50% with story-sentiment-visualiser).

   After participants interacted with an interface for 2 minutes, they were shown a keyword. Participants had to submit this keyword for payment to be honoured. This checked that participants tested the interface for the required time and that it loaded correctly. No participants were rejected. Training time was not included for the interfaces because the tools were designed to be intuitive and to evaluate RiCEv1 with open-ended CSTs where discovery and self-learning are key (Hewett, 2005; Shneiderman, 2002; Shneiderman et al., 2006).

3. **RiCEv1:** Considering the interface they had just used, participants scored the RiCEv1 items as described in Table 3.7. Taking direction from related questionnaires (Cherry & Latulipe, 2014; Grant et al., 2002; O'Brien et al., 2018), 5 mean averages are derived for Reflection on Current Process (RiCEv1-Cp), Reflection on Self (RiCEv1-Se), Reflection through Experimentation (RiCEv1-Ex), Reflection on Past Experiences (RiCEv1-Pa), and a total RiCEv1 score (see Table 3.7).

4. **CSI:** Participants completed the CSI (Cherry & Latulipe, 2014) for the interface they had just used to examine how RiCEv1 correlates with the CSI. This included completing both the CSI's item scoring and factor comparison sections (see Section 2.3.2.1). The weighted sum of the means is calculated for a total CSI score.
5. **Repeat:** Steps 2, 3 and 4 are repeated for the other interface.
6. **Comparison:** A comparison question was asked to test if RiCEv1 or its factors are higher for the interface where most participants agree they experienced the most moments of reflection. Participants were shown images of the two user interfaces and asked:

   > "When exploring the two interfaces [pictured], with which did you experience the most moments of reflection?"

Finally, participants could offer further comments via an open-ended text box. A week later, participants were invited to repeat the study procedure to assess RiCEv1's test-retest reliability.

4.1.4 Data Analysis Method

This section describes the statistical techniques used to test RiCEv1. The methodological rationale for the analysis follows standard practice in the evaluation of questionnaires (Boateng et al., 2018).

4.1.4.1 Confirmatory Factor Analysis

To test RiCEv1's factor structure, a Confirmatory Factor Analysis (CFA) was used on the data collected in the test and re-test conditions for both sound-sketcher and story-sentiment-visualiser. The lavaan package for the R programming language (Rosseel, 2012) is used to support reproducibility. Each pair of statements from RiCEv1 was modelled as loading onto their respective factor, as identified from the EFA (see Table 3.7). The maximum likelihood estimator with Satorra-Bentler scaling (robust maximum likelihood) was applied as Finney and DiStefano (2006) found this to be appropriate for ordinal data with 6 or more points.

Metrics of the CFA model's fit are examined as described by Kline (2015) and commonly used across HCI studies (Bowman et al., 2021; Cai et al., 2022; Makransky et al., 2017; Vahlo & Karhulahti, 2020). These metrics are (Boateng et al., 2018; Kline, 2015; Matsunaga, 2010): a *chi-squared test* to assess the difference between the sample's covariance and the model's covariance; *the Comparative Fit Index (CFI)* and *Tucker-Lewis Index (TLI)* to assess the ratio between the deviation of the model from the worst fitting, and best fitting, model; the *Root Mean Squared Error of Approximation (RMSEA)* to measure the degree of the model's misspecification; and the *Standardised Root Mean Square Residual (SRMR)* to assess the error between the model's covariance and the sample's covariance.

The acceptability of each metric is determined based on the following criteria: chi-squared test is not significant ($p \geq 0.05$) (Matsunaga, 2010); CFI and TLI $\geq .90$ is acceptable (Bentler & Bonett, 1980; Hair et al., 1995) and $\geq .95$ is excellent (Boateng et al., 2018; Kline, 2015); RMSEA ≤ 0.08 and not significant ($p \geq 0.05$) is acceptable (Matsunaga, 2010); and SRMR ≤ 0.08 is acceptable (Boateng et al., 2018; Hair et al., 1995; Matsunaga, 2010).

4.1.4.2 Test-Retest Reliability

Test-retest reliability is the extent to which people's questionnaire responses are the same between points in time. Following Boateng et al. (2018) and the TSRI (Bentvelzen et al., 2021), the Intra-Class Correlation (ICC) coefficient is calculated for RiCEv1's factors. Points are taken from the first survey responses and 1 week later for both interfaces. The results are interpreted using Koo and Mae's (2016) guidelines of poor (ICC $\leq .5$), moderate ($.5 <$ ICC $> .75$), good ($.75 \leq$ ICC $> .9$) and excellent (ICC $\leq .9$).

4.1.4.3 Differentiation by Known Groups

To evaluate how well RiCEv1 captures the intended measure, the difference between RiCEv1 and its factors is examined for both interfaces. Using the Wilcoxon signed-rank test (as the data is ordinal), the medians are compared for significantly different factors against the count of users who selected the interface they found they had the most moments of reflection with. This determines if the factors move in the same direction.

4.1.4.4 Comparison with Existing Scales

Correlations are identified between RiCEv1's total score and the total SRIS scores (Grant et al., 2002), and CSI (Cherry & Latulipe, 2014). This assesses that RiCEv1 captures the intended measure and not derivatives of these related scales. The intuition is that higher SRIS scores occur alongside higher RiCEv1 scores, and higher CSI scores occur alongside higher RiCEv1 scores. Weak ($\geq .3$ and $< .5$) to moderate ($\geq .5$ and $< .7$) correlations are expected to support the notion that, although RiCEv1 is conceptually different, it is still influenced by related factors. Given this, the following hypotheses are tested:

- **H1:** For story-sentiment-visualiser, there will be a weak to moderate positive correlation between RiCEv1's total score and the SRIS's total score.
- **H2:** For story-sentiment-visualiser, there will be a weak to moderate positive correlation between RiCEv1's total score and the CSI's total score.
- **H3:** For sound-sketcher, there will be a weak to moderate positive correlation between RiCEv1's total score and the SRIS's total score.
- **H4:** For sound-sketcher, there will be a weak to moderate positive correlation between RiCEv1's total score and the CSI's total score.

Only correlations between total scores are inspected instead of individual factors to focus on testing RiCEv1 as a whole and avoid family-wise type 1 errors on account of multiple tests. Spearman's (1904) Rho correlation coefficient is used as it is suited to ordinal data.

4.2 Results

This section reports the results of the statistical tests outlined above.

4.2.1 Confirmatory Factor Analysis

Table 4.2 shows the fit metrics for the CFA of RiCEv1. CFI is acceptable in both re-test conditions. SRMR is also acceptable in both test conditions and re-test conditions for sound-sketcher. There are five acceptable metrics for the re-test condition of story-sentiment-visualiser for the Chi-squared test and RMSEA. Other metrics do not achieve acceptance.

Table 4.2: Fit metrics for RiCEv1's confirmatory factor analysis across conditions and interfaces. Acceptable metrics are in bold.

Timing	Interface	Chi-squared	CFI	TLI	RMSEA	SRMR
	Criterion:	**$p \geq 0.05$**	**≥ 0.9**	**≥ 0.9**	**RMSEA ≤ 0.08; $p \geq 0.05$**	**≤ 0.08**
Test	Sound	$\chi^2(14) = 38.0$, p = < .001	0.88	0.75	RMESA = 0.17 90% CI [0.11, 0.24] $p < .001$	**0.07**
Re-test	Sound	$\chi^2(14) = 31.0$, $p < .001$	**0.91**	0.82	RMESA = 0.16 90% CI [**0.08**, 0.23] p = 0.01	**0.08**
Test	Story	$\chi^2(14) = 33.3$, $p < .001$	0.89	0.79	RMESA = 0.17 90% CI [0.10, 0.25] p = 0.01	0.09
Re-test	Story	$\chi^2(14) = 20.8$, **p = 0.11**	**0.94**	0.88	RMESA = 0.12, 90% CI [**0.00**, 0.21], **p = 0.15**	**0.07**

4.2.2 Test-Retest Reliability

Table 4.3 shows the ICCs between the test and re-test measures for RiCEv1 and its factors. For story-sentiment-visualiser, a moderate test-retest reliability is inferred for all factors. The confidence intervals range from poor to moderate, excluding RiCEv1-Ex, which shows poor test-retest reliability. ICCs for sound-sketcher also range from poor to moderate. Total RiCEv1 ICCs show moderate test-retest reliability for both interfaces.

Table 4.3: Intra-class correlations between the test and re-test measures for RiCEv1 and its factors. Significant measures in bold.

Interface	RiCEv1	ICC2	p	CI Lower	CI Upper
Story	RiCEv1-Ex	.22	.055	.13	.30
Story	**RiCEv1-Se**	**.52**	**< .001**	**.45**	**.59**
Story	**RiCEv1-Cp**	**.51**	**< .001**	**.44**	**.58**
Story	**RiCEv1-Pa**	**.51**	**< .001**	**.43**	**.57**
Story	**RiCEv1**	**.61**	**< .001**	**.55**	**.67**
Sound	**RiCEv1-Ex**	**.45**	**< .001**	**.37**	**.52**
Sound	**RiCEv1-Se**	**.64**	**< .001**	**.58**	**.69**
Sound	**RiCEv1-Cp**	**.43**	**< .001**	**.35**	**.50**
Sound	**RiCEv1-Pa**	**.47**	**< .001**	**.39**	**.54**
Sound	**RiCEv1**	**.58**	**< .001**	**.52**	**.64**

4.2.3 Differentiation by Known Groups

When completing the study for the first time, 60.3% of participants selected experiencing the most moments of reflection with story-sentiment-visualiser instead of sound-sketcher (39.7%). This pattern continued when participants completed the study 1 week later (64.8% story-sentiment-visualiser, 35.2% sound-sketcher).

Wilcoxon signed-rank tests were conducted for RiCEv1 and its factors, reported in Table 4.4. For both the test and re-test responses, RiCEv1-Ex scores were significantly *lower* for story-sentiment-visualiser than sound-sketcher. RiCEv1-Se scores were significantly *higher* for story-sentiment-visualiser than sound-sketcher.

Table 4.4: Wilcoxon signed-rank tests across the interfaces for RiCEv1 on both test and re-test. Significant results are in bold.

Timing	RiCEv1	V	p	Median for Story	Median for Sound
Test	**RiCEv1-Ex**	**501.0**	**.038**	**6.0**	**7.0**
Test	**RiCEv1-Se**	**974.0**	**.046**	**6.0**	**5.3**
Test	RiCEv1-Cp	673.5	.891	6.8	7.0
Test	RiCEv1-Pa	960.5	.111	7.5	6.0
Test	RiCEv1	828.0	.810	5.9	6.4
Re-test	**RiCEv1-Ex**	**166.0**	**< .001**	**5.8**	**7.5**
Re-test	**RiCEv1-Se**	**788.5**	**.002**	**6.0**	**4.5**
Re-test	RiCEv1-Cp	476.0	.483	7.0	6.5
Re-test	RiCEv1-Pa	685.5	.058	8.0	7.0
Re-test	RiCEv1	627.0	.693	6.2	6.2

4.2.4 Comparison with Existing Scales

Recall the hypotheses from Section 4.1.4.4. For story-sentiment-visualiser, there is a weak positive correlation between RiCEv1 and the SRIS score on test (r(58) = .36, p = .006) and re-test (r(54) = .40, p = .003) – accept H1. There is a moderate positive correlation between RiCEv1 and the CSI on test (r(58) = .52, p < .001) and re-test (r(54) = .66, p < 0.001) – accept H2. For sound-sketcher, there is a weak positive correlation between the RiCEv1 and SRIS scores on test (r(58) = .31, p = .018) and re-test (r(54) = .37, p = 0.006) – accept H3. Between RiCEv1 and the CSI, there is a moderate positive correlation on test (r(58) = .54, p < 0.001) and re-test (r(54) = .67, p < 0.001) – accept H4.

4.3 Discussion

Table 4.5: Summary of Chapter 4's main findings.

Main Finding	Location
RiCEv1 reliably measures a distinct construct, whilst also relating to scales on creativity and people's capacity for reflection.	§4.2.4
RiCEv1 reliably shows differences between different types of reflection for different interfaces. — Self-reflection is characterised as occurring more so in people's use of a CST with aspects of creative writing than a CST with aspects of drawing and music. — Reflection-through-experimentation is characterised as occurring more in people's use of a CST with aspects of drawing and music than a CST with aspects of creative writing.	§4.2.3
RiCEv1 shows stronger reliability for participants more familiar with the CST tested and the creative task.	§4.1.4.1 §4.1.4.2

This chapter evaluated RiCEv1's reliability. A user study applied RiCEv1 to two novel interfaces and creative tasks related to creative writing and music making. The main findings are summarised in Table 4.5. The factors show good content validity and successfully distinguish between the types of reflection common in different CST interactions. The findings are discussed below, and limitations are described in Section 4.3.1.

The results show that RiCEv1 measures moments of reflection and not a different construct. RiCEv1 correlated with the SRIS (Grant et al., 2002), showing that higher RiCEv1 scores occur alongside more naturally reflective people. RiCEv1 also correlated with the CSI (Cherry & Latulipe, 2014), showing that higher RiCEv1 scores occur alongside interfaces that better foster creativity. These correlations were moderate, demonstrating that, whilst related, RiCEv1 measured a distinct construct. When combined with the expert review process in the previous chapter (see Section 3.2.2), this demonstrates strong content validity.

The differentiation by known groups test shows that RiCEv1 differentiates between different *types* of reflection in story-sentiment-visualiser and sound-sketcher. The differences between RiCEv1-Se's and RiCEv1-Ex's medians show that moments of *self*-reflection (RiCEv1-Se) occurred more so with story-sentiment-visualiser. Moments of reflecting through experimentation (RiCEv1-Ex) occurred more with sound-sketcher. Bentvelzen et al. (2022) found that comparisons to an absolute reference encourage reflection. This corroborates the finding that more self-reflection occurred with story-sentiment-visualiser, which offers an evaluation metric for comparison through its smiley scale. In contrast, RiCEv1-Ex was higher for sound-sketcher because of its open-ended interaction. Participants continually evaluated their creations against their own criteria through experimentation. It is notable that differences between factors were found despite the participants only marginally scoring story-sentiment-visualiser as more reflective in Section 4.2.3. A study comparing interfaces with a stronger split of opinion will thus present more prominent differences between RiCEv1 and its factors.

RiCEv1 and its factors show moderate to poor test-retest reliability. The significant differences between RiCEv1-Ex and RiCEv1-Se for story-sentiment-visualiser and sound-sketcher show that test-retest reliability varies between creative disciplines, where different types of reflection are more or less prominent. The CFA fit is also better for story-sentiment-visualiser than sound-sketcher, demonstrating that RiCEv1 is task dependent.

RiCEv1's scores will also be different for different phases of the creative process, which is complex and non-linear (see Section 2.1). For example, as the tasks for the user interfaces tested were open-ended, it is difficult to deduce whether participants were in divergent or convergent thinking phases. This is examined further in a study that investigates measures of RiCE throughout a creative process (see Chapter 8).

Test-retest reliability would improve if participants investigated the interfaces for longer, or if longer than one week was left between data collection points. This mitigates for learning effects. The stronger fit of the CFA in the re-test conditions shows that RiCEv1 more reliably measures reflection when participants are familiar with a creative interface or task. Story-sentiment-visualiser and its associated task (writing a story) are also likely more familiar than sound-sketcher's task (making music from drawings), hence participants choosing it as more reflective.

4.3.1 Limitations

RiCEv1's reliability is limited to the assessments in this formative user study. The CFA results limit the reliability of RiCEv1. However, the CFA only indicates RiCEv1's fit because: the scale has two items per factor (Kline, 2015, pg. 201) whereas three or more items are recommended for CFA to have sufficient information (ibid); and "the sample size [is relatively speaking] not large" (Kline, 2015, pg. 259) for a CFA analysis. This limits the extent to which the CFA results can be interpreted. However, this thesis argues that the current CFA results show potential for future work and tend towards good fit. The key strength of RiCEv1's reliability is its content validity, which is achieved through the consultation with 10 professionals in creative disciplines and large-scale survey experiments with multiple CST users. The differentiation by known groups test also confirms that RiCEv1 differentiates between how much different types of reflection occur in different CSTs.

RiCEv1 needs to be extended to increase the number of items per factor to address its limitations. This can include adding reverse-scored items to ensure consistency in participants' scoring. Indeed, the lower reliability scores may have resulted from selecting only two items per factor. More items per factor would also allow for more refined designs of RiCEv1. Inspection of the correlations between items (Kline, 2015) would guide the design of alternative models for RiCE. The inclusion of extra items should be balanced against questionnaire length. RiCEv1 is intended to be used quickly alongside other measures to not increase participants' burden. Scales with comparable goals include between 9 and 12 items (Bentvelzen et al., 2021; Cherry & Latulipe, 2014; Jackson et al., 2008; O'Brien et al., 2018). The limitations are addressed to update RiCEv1 to RiCEv2 in the following chapter.

4.4 Conclusion

This chapter evaluated the reliability of RiCEv1. This was achieved through a user study of two CSTs with elements of creative writing, music and sketching. RiCEv1 was shown to have strong content validity and could reliably identify differences between types of reflection in different CST interactions. The following chapter further updates RiCEv1 to RiCEv2 to improve upon its reliability for use in the thesis case study domain.

Chapter 5

Design of RiCE Version 2

Chapter 3 showed that the Reflection in Creative Experience Questionnaire Version 1 (RiCEv1) has strong content validity and four factors which are interpretable in a conceptually meaningful way (Worthington & Whittaker, 2006). Chapter 4 showed that RiCEv1 can successfully differentiate between different types of reflection in different CSTs. This chapter updates RiCEv1 to improve its construct validity: how well its structural model fits other data. This improves upon RiCEv1 before the thesis moves to its case study domain of interest.

The chapter is organised as follows. Section 5.1 outlines the process used to iterate RiCEv1 to RiCEv2. Section 5.2 details the findings of this process. Section 5.3 closes the chapter by discussing RiCEv2's limitations.

5.1 Method

To remind the reader, two limitations of RiCEv1 identified in Chapter 4 were that there are too few items per factor to be able to assess construct validity (Kline, 2015) and no items with reverse wording to check for consistency in participants' scoring (Müller et al., 2014). A process based on Kline (2015) is followed to address these limitations. First, RiCEv1 is extended by revisiting the results from the exploratory factor analysis (Section 3.2) to add more items per factor. Correlation and inter-rater reliability metrics are then inspected iteratively, removing items to reduce the extended questionnaire until acceptable metrics are produced. The process is applied to a dataset from the user study presented in Chapter 11, as described below.

5.1.1 Dataset

The dataset from the user study presented in Chapter 11 is used to refine an extended RiCEv1 questionnaire to RiCEv2. The user study applied the extended questionnaire to evaluate people's interaction with a novel CST for music composition with AIGC, named wAIve. wAIve allows users to write short music compositions using AIGC components and animations to encourage reflection. 22 participants were recruited who are music novices and study computer science-related subjects. They started with the same level of familiarity with wAIve for consistency and were tasked to compose a piece of music for twenty minutes. Full details on participants' demographics, expertise and the study task are in Chapter 11.

The decision to use this data was because it is representative of a typical CST user study (Frich et al., 2019). The data also directly relates to the thesis case study domain, where RiCEv2 is later applied. This thus supports the reliability of the RiCEv2 findings presented in the later chapters of this thesis.

5.1.2 Data Analysis Method

The dataset was split into a training set (14 out of 22 participants) and test set (8 out of 22 participants). Over several iterations, items were removed based on inspection of correlation and inter-item reliability metrics, as follows:

Convergent Validity: Items were identified with negative correlations to other items within a factor. This shows that items do not measure the same underlying construct. Positive correlations for items within a factor were retained, whilst negatively correlated items were removed. Spearman's (1904) Rho was used to assess the correlations between items in a factor as it is suited to ordinal, non-normal data. The following guidelines are used to interpret the correlations: poor (< .3), weak (>= .3 and < .5), moderate (>= .5 and < .7) and strong (>= .7).

Inter-item Reliability: The inter-item reliability is assessed to examine whether scoring is consistent for items within factors and the whole scale. Cronbach's (1951) alpha is used as it is suitable for non-binary ordinal data.

Following general guidelines, values >= .7 are found as acceptable, whilst being lenient because scales with few items per construct naturally yield lower alphas (Schrepp, 2020) and RiCEv2 aims to be lightweight.

The metrics are assessed using the training set data, and items are removed until acceptable metrics are found. The metrics are then inspected for the test set to evaluate RiCEv2's generalisability to similar data.

5.2 Design Iterations

This section first describes how RiCEv1 was extended to prepare for its refinement. This is followed by descriptions of the changes made through each iteration of the data analysis method, described above. The final RiCEv2 design is then reported.

5.2.1 Extending RiCEv1

RiCEv1 was first extended by revisiting the results from the exploratory factor analysis in Section 3.2. This extended questionnaire used all items with acceptable loadings for each factor; RiCEv1 initially only used the top two. Thus, there were more items per factor. This would enable a confirmatory factor analysis in future (Kline, 2015).

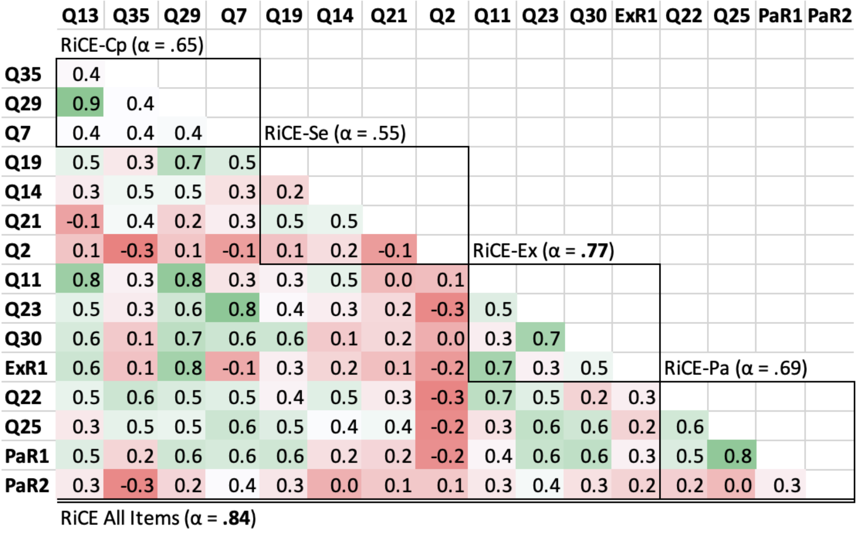
Two new items were created by reversing the wording of the highest loading items. The reverse wording would help to assess a participant's scoring consistency. This also led to an equal number of items per factor, described as best practice in Müller et al. (2014). This led to a questionnaire with four items per factor, shown in Table 5.1.

5.2.2 Training Set: Iteration 1

Figure 5.1 shows the correlation matrix and alpha values for the training set data, for the extended RiCEv1 questionnaire. Three items were removed based on this matrix. These were: Q2, as it had only weak and negative correlations with other items; PaR2, with mostly weak or negative correlations; and Q35, with the weakest correlations to other items compared within the RiCE-Cp factor.

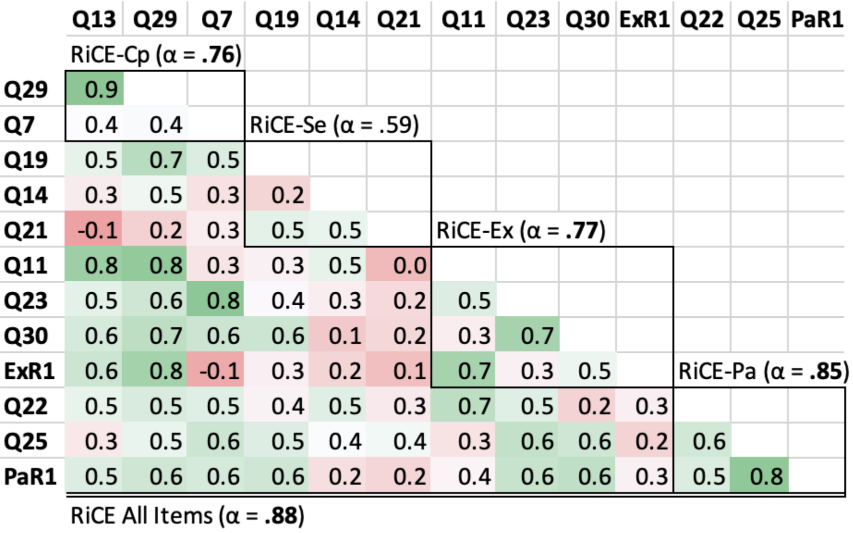
Table 5.1: The extensions to the RiCEv1 questionnaire. These are later refined to create RiCEv2. Question numbering is from the original question list for RiCEv1's design in Table 3.3.

The Extended RiCEv1
Considering your recent experience of [SYSTEM], please indicate the extent to which you agree with the following statements:
Factor: Reflection on Current Process
Q13: Whilst being creative, I liked to think about my actions to find alternative ways of doing them.
Q35: I often re-examined things I'd already learnt.
Q29: I considered different ways of doing things.
Q7: I found myself iteratively refining and assessing my creative process.
Factor: Reflection on Self
Q19: I learned many new things about myself during the experience.
Q14: I pondered over the meaning of what I was doing in relation to my personal experience.
Q21: I often reappraised my experiences with the system so I could learn from them.
Q2: I considered how my outputs from the system might be interpreted differently in the future.
Factor: Reflection through Experimentation
Q11: I made comparisons within the system to consider alternative ways of doing things.
Q23: I often generated, tested and revised ideas.
Q30: I often reflected on my actions to see whether I could have improved on what I did.
ExR1: I made no comparisons within the system to consider alternative ways of doing things.
Factor: Reflection on Past Experiences
Q22: I explored my past experiences as a way of understanding new ideas.
Q25: Whilst creating, I thought back on some of my past experiences.
PaR1: Whilst creating, I did not think about my past experiences.
PaR2: I never explored my past experiences to understand new ideas.



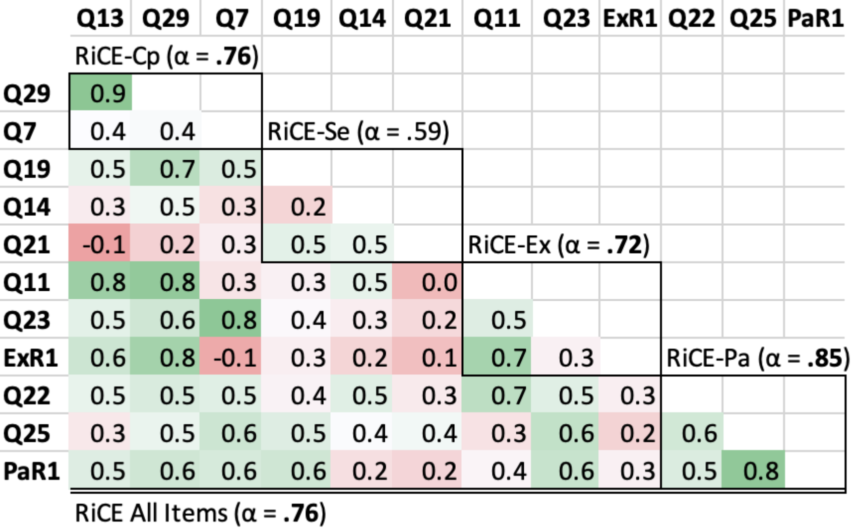
	Q13	Q35	Q29	Q7	Q19	Q14	Q21	Q2	Q11	Q23	Q30	ExR1	Q22	Q25	PaR1	PaR2
	RiCE-Cp (α = .65)															
Q35	0.4															
Q29	0.9	0.4														
Q7	0.4	0.4	0.4		RiCE-Se (α = .55)											
Q19	0.5	0.3	0.7	0.5												
Q14	0.3	0.5	0.5	0.3	0.2											
Q21	-0.1	0.4	0.2	0.3	0.5	0.5										
Q2	0.1	-0.3	0.1	-0.1	0.1	0.2	-0.1		RiCE-Ex (α = **.77**)							
Q11	0.8	0.3	0.8	0.3	0.3	0.5	0.0	0.1								
Q23	0.5	0.3	0.6	0.8	0.4	0.3	0.2	-0.3	0.5							
Q30	0.6	0.1	0.7	0.6	0.6	0.1	0.2	0.0	0.3	0.7						
ExR1	0.6	0.1	0.8	-0.1	0.3	0.2	0.1	-0.2	0.7	0.3	0.5		RiCE-Pa (α = .69)			
Q22	0.5	0.6	0.5	0.5	0.4	0.5	0.3	-0.3	0.7	0.5	0.2	0.3				
Q25	0.3	0.5	0.5	0.6	0.5	0.4	0.4	-0.2	0.3	0.6	0.6	0.2	0.6			
PaR1	0.5	0.2	0.6	0.6	0.6	0.2	0.2	-0.2	0.4	0.6	0.6	0.3	0.5	0.8		
PaR2	0.3	-0.3	0.2	0.4	0.3	0.0	0.1	0.1	0.3	0.4	0.3	0.2	0.2	0.0	0.3	
	RiCE All Items (α = **.84**)															

Figure 5.1: Iteration 1's training set correlations and alpha values for RiCEv2's design.



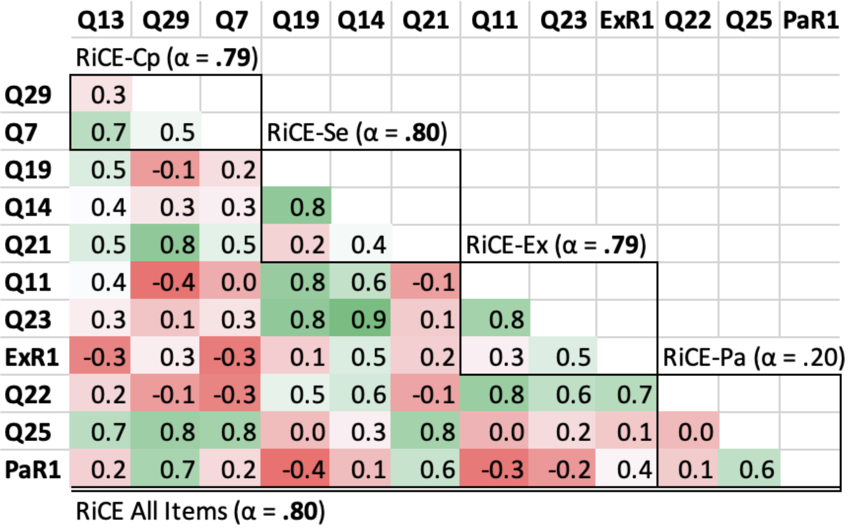
	Q13	Q29	Q7	Q19	Q14	Q21	Q11	Q23	Q30	ExR1	Q22	Q25	PaR1
	RiCE-Cp (α = **.76**)												
Q29	0.9												
Q7	0.4	0.4		RiCE-Se (α = .59)									
Q19	0.5	0.7	0.5										
Q14	0.3	0.5	0.3	0.2									
Q21	-0.1	0.2	0.3	0.5	0.5		RiCE-Ex (α = **.77**)						
Q11	0.8	0.8	0.3	0.3	0.5	0.0							
Q23	0.5	0.6	0.8	0.4	0.3	0.2	0.5						
Q30	0.6	0.7	0.6	0.6	0.1	0.2	0.3	0.7					
ExR1	0.6	0.8	-0.1	0.3	0.2	0.1	0.7	0.3	0.5		RiCE-Pa (α = **.85**)		
Q22	0.5	0.5	0.5	0.4	0.5	0.3	0.7	0.5	0.2	0.3			
Q25	0.3	0.5	0.6	0.5	0.4	0.4	0.3	0.6	0.6	0.2	0.6		
PaR1	0.5	0.6	0.6	0.6	0.2	0.2	0.4	0.6	0.6	0.3	0.5	0.8	
	RiCE All Items (α = **.88**)												

Figure 5.2: Iteration 2's training set correlations and alpha values for RiCEv2's design.



	Q13	Q29	Q7	Q19	Q14	Q21	Q11	Q23	ExR1	Q22	Q25	PaR1
	RiCE-Cp (α = .76)											
Q29	0.9											
Q7	0.4	0.4		RiCE-Se (α = .59)								
Q19	0.5	0.7	0.5									
Q14	0.3	0.5	0.3	0.2								
Q21	-0.1	0.2	0.3	0.5	0.5		RiCE-Ex (α = .72)					
Q11	0.8	0.8	0.3	0.3	0.5	0.0						
Q23	0.5	0.6	0.8	0.4	0.3	0.2	0.5					
ExR1	0.6	0.8	-0.1	0.3	0.2	0.1	0.7	0.3		RiCE-Pa (α = .85)		
Q22	0.5	0.5	0.5	0.4	0.5	0.3	0.7	0.5	0.3			
Q25	0.3	0.5	0.6	0.5	0.4	0.4	0.3	0.6	0.2	0.6		
PaR1	0.5	0.6	0.6	0.6	0.2	0.2	0.4	0.6	0.3	0.5	0.8	
	RiCE All Items (α = .76)											

Figure 5.3: Iteration 3's training set correlations and alpha values for RiCEv2's design.



	Q13	Q29	Q7	Q19	Q14	Q21	Q11	Q23	ExR1	Q22	Q25	PaR1
	RiCE-Cp (α = .79)											
Q29	0.3											
Q7	0.7	0.5		RiCE-Se (α = .80)								
Q19	0.5	-0.1	0.2									
Q14	0.4	0.3	0.3	0.8								
Q21	0.5	0.8	0.5	0.2	0.4		RiCE-Ex (α = .79)					
Q11	0.4	-0.4	0.0	0.8	0.6	-0.1						
Q23	0.3	0.1	0.3	0.8	0.9	0.1	0.8					
ExR1	-0.3	0.3	-0.3	0.1	0.5	0.2	0.3	0.5		RiCE-Pa (α = .20)		
Q22	0.2	-0.1	-0.3	0.5	0.6	-0.1	0.8	0.6	0.7			
Q25	0.7	0.8	0.8	0.0	0.3	0.8	0.0	0.2	0.1	0.0		
PaR1	0.2	0.7	0.2	-0.4	0.1	0.6	-0.3	-0.2	0.4	0.1	0.6	
	RiCE All Items (α = .80)											

Figure 5.4: Iteration 4's test set correlations and alpha values for RiCEv2's design.

5.2.3 Training Set: Iteration 2

Figure 5.2 shows the correlation matrix and alpha values for the training set data with the items above removed. All alpha values were found as acceptable, except for RiCE-Se. However, the decision was made not to remove an item from RiCE-Se because a minimum of three items per factor is recommended for a confirmatory factor analysis (Kline, 2015). Instead, Q30 was removed from the RiCE-Ex factor to support RiCE's future usability – three items for all factors make the scale more consistent and simpler for HCI researchers to adopt and use.

5.2.4 Training Set: Iteration 3

Figure 5.3 shows the correlation matrix and alpha values with three items per factor. Although smaller alpha values were found overall compared to the previous iteration, they were still acceptable. The future ease of having three items per factor was prioritised.

5.2.5 Test Set: Iteration 4 and RiCEv2 Final Design

The analysis for the new RiCE structure with three items per factor was repeated on the test set data. Figure 5.4 shows its correlation matrix and alpha values. All items within the factors are positive. This shows construct validity. However, RiCE-Pa has poor inter-item reliability and weak correlations. The decision was made to remove the RiCE-Pa factor. Thus, the final design of RiCEv2 is shown in Table 5.2.

5.3 Discussion

This chapter documented the development of RiCEv2. RiCEv2 is presented in Table 5.2. It was iteratively refined based on data collected in Chapter 11. It includes more items per factor than RiCEv1 and a reversed item to support its reliability. It includes factors of reflection-on-process (RiCEv2-Cp), reflection-on-self (RiCEv2-Se) and reflection-through-experimentation (RiCEv2-Ex); reflection-on-past was removed as it showed poor inter-rater reliability in the test set data.

A large-scale confirmatory factor analysis is needed to demonstrate that RiCEv2's factors are conceptually distinct and do not overlap (Kline, 2015).

Table 5.2: Items and instructions for administering and scoring RiCEv2.

RiCEv2: VERSION 2

::: INSTRUCTIONS FOR ADMINISTERING :::

When administering RiCEv2, each item should be placed along an 11-point scale from “Highly Disagree” (left) to “Highly Agree” (right). Values for each item are zero indexed, i.e., integers from 0 to 10. Please follow the question wording exactly, replacing only the name of your system where indicated. Dimension identifiers (e.g. Cp1), descriptions, and headings should not be visible to participants. Item order should be randomised.

Considering your recent experience of [SYSTEM], please indicate the extent to which you agree with the following statements:

Factor: Reflection on Current Process (RiCEv2-Cp)

Cp1 (Q13): Whilst being creative, I liked to think about my actions to find alternative ways of doing them.
Cp2 (Q29): I considered different ways of doing things.
Cp3 (Q7): I found myself iteratively refining and assessing my creative process.

Factor: Reflection on Self (RiCEv2-Se)

Se1 (Q19): I learned many new things about myself during the experience.
Se2 (Q14): I pondered over the meaning of what I was doing in relation to my personal experience.
Se3 (Q21): I often reappraised my experiences with the system so I could learn. from them.

Factor: Reflection through Experimentation (RiCEv2-Ex)

Ex1 (Q11): I made comparisons within the system to consider alternative ways of doing things.
Ex2 (Q23): I often generated, tested and revised ideas.
Ex3 (ExR1): I made no comparisons within the system to consider alternative ways of doing things. **Note: Reverse scoring.**

::: INSTRUCTIONS FOR SCORING:::

Following the design of related questionnaires (Cherry & Latulipe, 2014; Grant et al., 2002; O’Brien et al., 2018), the total RiCEv2 score (out of 10) is calculated as (Cp1+Cp2+Cp3+Se1+Se2+Se3+Ex1+Ex2+Ex3) ÷ 9. Each of the 3 factors are calculated as the average of its items. For example, Reflection on Current Process is (Cp1+Cp2+Cp3) ÷ 3.

This is challenging in creative contexts where reflection is abstract and unlikely to be easily separated into distinct categories. However, this thesis will show throughout the following chapters that RiCEv2 can be applied to assess if more or less of different types of reflection occour different different CST contexts. Its theoretical basis is also grounded by the systematic research reported in the previous two chapters.

As RiCEv2 was refined based on the user study in Chapter 11, its assessment is based in the context of AI-based music composition. This demonstrates its reliability within the study domain area of this thesis. Future research is needed to assess the potential for RiCEv2 to apply to other domains.

5.4 Conclusion

This chapter improved upon the limitations of RiCEv1. This was achieved by extending RiCEv1 and refining its design based on inspection of correlation matrices and inter-rater reliability metrics, using a dataset related to the thesis case study domain. This led to the design of RiCEv2 with three items per the three factors of reflection-on-process, reflection-on-self, and reflection-through-experimentation. With a tool for systematically assessing reflection in CST interaction in place, the thesis now moves to its case study on AI-based music composition.

Part II

Reflection in AI-based Music Composition Tools

Chapter 6

State-of-the-Art: AI-based Music Composition

The previous chapters showed that there are few Creativity Support Tools (CSTs) which focus on reflection, and no questionnaire approach for evaluating people's interaction with a CST. This motivated the development of the Reflection in Creative Experience Questionnaire (RiCEv2). The focus now moves to the thesis's case study domain of Artificial Intelligence (AI)-based music composition.

This chapter reviews the state-of-the-art in AI music making. It shows that there are few studies on how people reflect using AI Generated Content (AIGC) in CSTs, and none that systematically evaluate reflection in music composition. This justifies the evaluation of reflection across AI music tools in the following chapter.

The chapter is organised as follows. First, the chapter introduces two central aspects of the domain of AI-based music composition: Section 6.1 introduces the music composition process; Section 6.2 summarises state-of-the-art generative AI models for music. These two aspects are then considered jointly through a review of CSTs which implement AIGC for people's interaction (Section 6.3), including for the music composition context. The chapter concludes with a summary of the critiques of literature.

6.1 Music Composition Process

Music composition is a fundamental form of human creativity (Freeman, 1998). Composition is not unique to music; it refers to where materials are constructed to create something new (Whittall, 2011). For music, the material used for construction is sound. Sound is manipulated in terms of its pitch, timbre (such as harmony and texture) and time (such as rhythm and structure).

Whittall (2011) defines music composition as "both the activity of composing and the result of composing". Blum (2001) similarly defines music composition as "the activity or process of creating music, and the product of such activity". These definitions are broad and offer no insights into what the composition process entails.

Burnard (2000) presents a definition of music composition that combines improvisation and composition. However, improvisation is distinct in that "the decisive aspects of composition occur during performance" (Blum, 2001). Thus, performers reflect differently when improvising. Music composition also overlaps with music production and dissemination (Vanka et al., 2023). For example, musicians sometimes produce music whilst composing in a Digital Audio Workstation[1] (DAW). Recall that this thesis focuses on the music composition *process*, whilst noting overlap between these activities.

The music composition process also varies across genres. Rock bands often construct and test ideas through jamming (Biasutti, 2012). Live coders[2] make edits on the fly and build patterns (McLean & Wiggins, 2010). Orchestral composers are noted to develop themes or motifs (McAdams, 2004).

Models of the music composition process give a general view of how people create music over time. This includes extensions to the creative process models in Section 2.1. In comparison to the creative process models, music composition models emphasise iterative movements between phases and the preparation stage (Sloboda, 1985; Webster, 2002; Younker, 2000). Psychology studies on digital music composition have also used interviews (Bennett, 1976) or data logging techniques (Collins, 2007; Nilsson & Folkestad,

[1] A Digital Audio Workstation is software used for professional music making.

[2] Referring to the live coding of music, where code is written and interpreted in real-time which outputs audio and music.

2005) to examine composers' decisions and develop music composition process models. This includes for children's composition processes (Burnard, 1995; Kennedy, 2002; Nilsson & Folkestad, 2005; Swanwick & Tillman, 1986; Younker, 2000).

Whilst models of the composition process define broad stages, none indicate the different types of reflection that occur. Based on selected music literature, common moments of the composition process that affect reflection are proposed below.

Preparation: People often create a scaffold for their music before starting to compose (Nash, 2011; Wallas, 1926; Webster, 2002). For instance, a composer might select instruments to use in their piece before ideating musical fragments (Folkestad et al., 1998). Here, composers reflect-for-action (Candy, 2019).

Auditioning: Small-scale editing of notes in music composition is common, and composers spend time experimenting with musical fragments (Collins, 2007; Folkestad et al., 1998). This relates to moments where people listen to music fragments after editing. Expert composers tend to spend less time editing their music between episodes of playback (Nash, 2011; Nash & Blackwell, 2012) – performing edits during, and reacting to, auditory feedback (Addessi et al., 2015; Nash, 2011; Nash & Blackwell, 2012). Through listening, composers are reflecting-in-action (Schön, 1983).

Contemplation: Long episodes of listening indicate contemplative flow states (Bryan-Kinns & Hamilton, 2012). These episodes are described as a distinct phase of the composition process (Younker, 2000). Collins (2007) names these macro events: where composers sit back to consider the piece as a whole, cf. reflection-at-a-distance (Candy, 2019). Here, composers are concerned more with their music's general feel, not nuanced details.

Vision: People often have broad ideas in their head of how their music should sound before expressing them by playing an instrument (Younker, 2000) or committing to notation (Sloboda, 1985; Webster, 2002). Similar to preparation stages, composers reflect to develop their vision before action, cf. reflection-for-action (Candy, 2019).

Divergence: As composers develop expertise, they move from understanding musical conventions to breaking these conventions (Swanwick & Tillman, 1986; Webster, 2002) cf. divergent thinking. Swanwick and Tillman (1986) consider reflecting on the tension between breaking convention and keeping within a particular musical style to be a prerequisite of a fully-developed, self-aware composer, cf. Boden's (1991) transformational creativity.

Across the models of the composition process, there are common phases. These include: having an initial idea, sketching out the idea, building the idea into a first draft, elaborating and refining the work, and working on later revisions (Bennett, 1976; Webster, 2002). With this understanding of the music composition process, the following section introduces the second component of the thesis case study domain: AI for music.

6.2 Generative AI Architectures for Music

Alongside the emergence of a new wave of AI research (Xu, 2019) are several advances in modelling the music composition process (Carnovalini & Rodà, 2020). Lovelace (1843) speculated that computers could generate music in the 1800s. Developments such as deep learning have now led to AI systems that produce convincing, high-quality musical outputs (Carnovalini & Rodà, 2020). This includes both symbolic representations of music such as MIDI[3], and raw audio (see Caillon and Esling (2021)).

Below is an overview of AI architectures for music. The type of AI this thesis focuses on is generative AI systems, where models are built from datasets to produce new data with similar statistical properties (Carnovalini & Rodà, 2020). The architectures surveyed were selected to provide an understanding of the types of generative AI systems that are used in music interaction contexts. Briot et al. (2017), Carnovalini and Rodà (2020) and Herremans et al. (2017) present broader overviews.

Rule-based: The earliest generative music systems were rule-based. For example, in Mozart's Dice Game, fragments of music are randomly com-

[3] MIDI (Musical Instrument Digital Interface) is the specification of a communications scheme that sends data points to control digital music devices. The data uses an event-based format to represent musical control (for example, note-on and note-off messages), in contrast to sample based formats (Loy, 1985).

bined based on the outcome of a dice roll (Carnovalini & Rodà, 2020). The Illiac Suite (Hiller & Isaacson, 1957) used rule-based grammars to control randomness.

The limitation of rule-based methods is that programmers must decide which rules to encode. This is challenging when musical rules are not formally defined (Carnovalini & Rodà, 2020). However, rules can be used to extend small datasets or curate the output of more complex models. Rule-based systems are also easier to examine in a user interface than black-box approaches. This interpretability allows composers to reflect on and interact with the rules.

Markov Chains: Markov Chains model a probabilistic sequence of events, where events could be music data such as chords or notes (Carnovalini & Rodà, 2020). Markov Chains' strength is in modelling small datasets. However, they struggle to model long-term musical variations (Roberts, Engel, et al., 2018).

A successful example of a Markov Chain system for music is the Continuator (Pachet, 2003). The Continuator uses a Markov Chain to generate musical responses to a person's improvisation. Musical phrases similar to those input by a person are played back in real-time. By producing phrases in real-time interaction, there is opportunity for reflection-in-action (Schön, 1983).

Neural Network: Neural network approaches overcome the limitation that Markov Chains cannot model long-term variation. Recurrent Neural Networks (RNNs) retain information from previous time steps via hidden states (Karpathy, 2015; Roberts, Engel, et al., 2018). However, as RNNs derive gradients based on prior calculations during backpropagation, they suffer from the vanishing gradient problem: multiplying values repeatedly by small decimal values drags them close to zero.

Long-Short Term Memory Networks (LSTMs) are a type of RNN extended to forget, update or output data via vector masks and normalising functions. This avoids vanishing gradient issues. Gated Recurrent Units (GRUs) similarly extend RNNs to avoid vanishing gradients but use only reset and update gates (Phi, 2018).

LSTMs and GRUs have been used in bi-directional architectures (Briot et al., 2017; Hadjeres et al., 2017; Pati et al., 2019) and stacked on top of one another (Eck & Schmidhuber, 2002; Sturm et al., 2016). This approach has created successful music generation tools, including Sturm et al.'s (2016) FolkRNN for folk music, and Eck and Schmidhuber's (2002) model for blues music. However, these models produce outputs without providing interactivity, limiting composers' opportunities to reflect-in-the-moment (Candy, 2019). Often, only an AI's temperature control, which determines its output's randomness, is exposed (Bryan-Kinns et al., 2021).

Transformers: Transformer models use a different approach to solving the vanishing gradient problem and retain long-term musical structure (Vaswani et al., 2017). They use attention mechanisms to relate parts of sequences to other parts, not the previous input. This enables them to capture relationships between musical phrases, even if they are far apart in the sequence. They can also be parallelised, increasing the computing speed.

Interest in transformers has accelerated since tools such as ChatGPT (OpenAI, 2022) have become mainstream. Before its popularity, Banar and Colton (2021) had manipulated a General Purpose Transformer (GPT) model to create extreme musical passages. They systematically evaluated how fine-tuning these models to varying degrees affects musical qualities in their output (Banar & Colton, 2022). Banar and Colton (2021) also used outputs from a GPT model within a user interface, loading outputs based on their match to musical metrics. However, this approach only allows users to reflect on an AI's output instead of the AI's internal workings. The control of musical metrics also bears little relationship to composers' broader artistic goals and intuitions (Kaschub & Smith, 2009; Whittall, 2011).

Variational Auto-Encoders: The deep learning architectures described above are constrained to a specific genre based on the training dataset (Briot et al., 2017). Variational Auto-Encoders (VAEs) use neural networks to encode a dataset into a smaller compressed *latent* representation, in turn decoded by another set of neural networks (Kingma & Welling, 2013) (see Figure 6.1). This allows the internal model to be modified and parsed to the decoder (Fabius & Van Amersfoort, 2014; Roberts, Engel, et al., 2018), to move between genres. For example, Fabius and Van Amersfoort (2014) used a VAE to capture different styles of video game music.

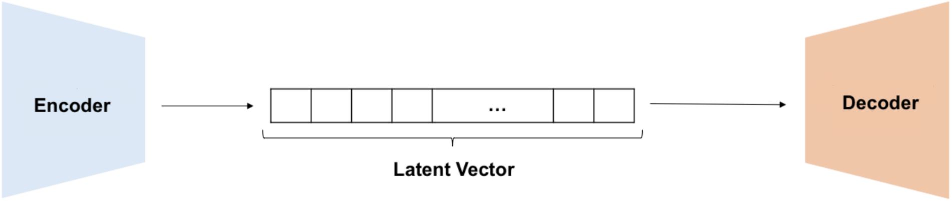


Figure 6.1: Visualisation of the variational auto-encoder architecture. Image adapted from Bryan-Kinns et al. (2021).

Users can control the internal model of a VAE. This provides unique opportunities for users to reflect on manipulations of this internal model using novel mappings. For example, Sonified Body (Murray-Browne & Tigas, 2021) uses a dancer's movement to alter values in the latent space of a generative music model. Several examples from Google Magenta[4] also allow users to make music by manipulating a VAE's latent vector. Notably, these interfaces used the MusicVAE model (Roberts, Engel, et al., 2018), which produces high-quality music that is steerable towards different genres by modifying the latent vector. An interaction design challenge here is in how to visualise the multidimensional latent vector representation within a graphical user interface. There is little research on how these mappings affect people's interaction in music composition, and none with a focus on reflection.

In summary, there are several critiques of AI architectures for music composition that are relevant to reflection. Rule-based systems and Markov Chains are interpretable, providing opportunity for reflection on their inner workings. However, they are limited to producing short musical phrases. LSTM networks generate more coherent output but are black-boxes; users must reflect on their outputs post-hoc. This restricts more reactionary reflection-in-action (Schön, 1983). VAEs' latent spaces can be mapped in unique ways to provide opportunity for reflection-in-the-moment; however, the interpretability of these mappings and how they support reflection remains uncertain. To better understand how AI affects reflection in interaction, the following section turns to their use in CSTs, including CSTs for music composition.

[4] https://magenta.tensorflow.org/demos/

6.3 AI-based Creativity Support Tools

Before the current crescendo of interest in generative AI, CST research was already exploring ways to enhance human creativity with technology. This included using computers to automate menial tasks up to the development of fully collaborative digital partners (Cornock & Edmonds, 1973; Lubart, 2005). Recent advances in AI have led to AIGC being used in CSTs to act more like collaborative partners, adding AI generated media indistinguishable from human creations to users' work. There are over 50 documented AI-based CSTs (Spoto & Oleynik, 2018) where AIGC contributes to a shared product with the user (Rezwana & Maher, 2022).

CSTs that allow humans to use AIGC have been called co-creative AI (Davis, 2021). This term implies that an AI, like a person, is also creative. Whether AI is creative or not is widely debated (Colton et al., 2020; Hertzmann, 2018; Shneiderman, 2022). Computational creativity researchers have argued how an AI might be perceived as creative (Colton & Wiggins, 2012), particularly if generating content which seems authentic to that which an AI might create based on its own experiences (Colton et al., 2018, 2020; Guckelsberger et al., 2017), or by generating insights on its internal processes (Charnley et al., 2012; Colton & Ventura, 2014).

Human-centred AI narratives (Garibay et al., 2023; Shneiderman, 2022) instead emphasise the view of AI as a tool which, although not creative in and of itself, empowers human creativity. Boden and Edmonds (2009) use the term computer-assisted art to describe the output of these tools because the computer acts as an aid in the users' creative process. This thesis is situated within HCI and aligns most closely with these human-centred narratives.

HCI researchers have also used the term mixed-initiative systems (Deterding et al., 2017) to describe CSTs with AIGC. This thesis does not use this term because, although AI tools might automatically add to a creative product, this is pre-programmed and does not demonstrate *initiative* in the cognitive sense that the word connotes. Instead, the term **AI-based CST** is used to be more neutral and not imply that an AI has initiative.

Casual Creators (Compton & Mateas, 2015; Compton, 2019) are a type of AI-based CST. They emphasise using AIGC to support non-professionals' initial, short-term enjoyment. The user experience is similar to Costello and Edmonds's (2007) pleasure of creation (see Section 2.1.2). Compton and Mateas (2015) present design patterns for Casual Creators. These patterns include to provide instant feedback and entertaining evaluations of people's creations. However, the design objectives of Casual Creators are tangential to the aim of supporting reflective interaction, emphasising fast, autotelic engagement rather than thoughtful and reflective interaction. Casual Creators also contrast human-centred AI narratives, which advocate for increased automation and control (Shneiderman, 2022); Casual Creators advocate for using AI to automate creation and decrease control.

6.3.1 AI-based CSTs for Reflection

There is little research on AI-based CSTs that is directly focused on reflection. An exception is Kreminski and Mateas's (2021) Reflective Creators. Reflective Creators are a subset of Casual Creators (Compton & Mateas, 2015). They aim to elicit reflection in casual, autotelic creative experiences, typically using AIGC. Some key design patterns for reflective creators are: reifying intent (ask users to make their intent explicit and provide mechanisms to negotiate intent), and interpretive refraction (to deliberately create distance between the user and their creation). However, it is difficult to show intent when creative intentions are indescribable, based on feelings or intuitions (Kaschub & Smith, 2009; Whittall, 2011). As a subset of Casual Creators, the premise of Reflective Creators is also not to spark reflection per se, but to create a reflective environment.

In light of limited research on reflection in AI-based CSTs, examples of how artists use AIGC relevant to reflection outside the music domain are briefly introduced. Caramiaux and Fdili Alaoui (2022) found that pioneering creators of AI artworks leverage the ambiguity of AI outputs by making glitches central to their process. This ambiguity introduced by AI can provide opportunities for reflection (Ford & Bryan-Kinns, 2022b; Wilson et al., 2023), such as to reflect on surprises (Candy, 2019) in the AI output. Yurman and Reddy (2022) used image-generating AI tools (Goodfellow et al., 2014) in their watercolour practice, finding that they needed to reflect on their own perspectives to assign meanings to ambiguous AIGC. Lewis (2023, 2025) found that ChatGPT (OpenAI, 2022), when acting like an art teacher, would provide suggestions influencing their drawing style; this sparked their reflection on the ownership of the data used by ChatGPT. How much agency an AI is given over a creative product is an issue commonly raised in human-AI interaction research (Amershi et al., 2019; Boden & Edmonds, 2009; Lewis, 2023; Louie et al., 2020; Wilson et al., 2023; Xambó, 2022).

Outside of the CST field, AIGC has been used in HCI to prompt consideration of different perspectives (Bentvelzen et al., 2022). Reicherts et al. (2022) compared an AI's text-based and voice-based design, which offers prompts to support human-human collaboration when identifying graph trends. When the AI was perceived as proactive (like a tutor), people were encouraged to think for themselves. When the AI was perceived as reactive (like an assistant), people let the AI "do the thinking" on their behalf. Similarly, Wagener et al. (2023) used voice-based AI prompts to support reflection when drawing art in virtual reality. They recommended not making prompts too specific to retain users' feelings of autonomy. Hubbard et al. (2021) tested AI prompts to scaffold children's storytelling, played as audio from their own stuffed toy to make a personal connection. Li et al. (2023) took this further, fully involving young girls in designing the identities of their own AI robot.

The examples of HCI research that have used AI outside of CST research above have prompted and motivated reflection during interaction – a limitation of the reflective CSTs discussed in Section 2.2.1. However, in these examples, AIGC interrupts creative flow (Csíkszentmihályi, 1990). This is problematic for CST interaction, especially when reflection occurs in-the-making-moment (Candy, 2019).

6.3.2 AI-based CSTs for Music Composition

This section brings together the music and AI components from the previous sections. There is a rich history of research on how people use digital interfaces for music making. This is exemplified by publication venues such as the New Instruments for Musical Expression conference (Poupyrev et al., 2001). However, human-centred research has only recently started to emerge on how people use AIGC in music making (Jourdan & Caramiaux, 2023). AI music research has focused mostly on modelling musical aspects (Herremans et al., 2017; Huang et al., 2020), and most AI tools afford limited interactivity (Bryan-Kinns et al., 2023). For example, generative AI systems have modelled actions in the composition process of connecting musical phrases (Pati et al., 2019) and adding harmonies (Huang & Chew, 2005; Louie et al., 2020). Below, the few examples of how AI generated music has been used in CSTs for music making are reviewed.

Generative AI has been used in music making to generate, curate and rearrange its outputs (Huang et al., 2020). Huang et al. (2020) found that teams of musicians and developers participating in the international AI songwriting contest[5] generated vast quantities of AIGC for later curation, rather than modifying AI models. Given this, CSTs have been designed to make curating and organising AIGC more manageable. For example, the Interactive Generative Music Environment (IGME), shown in Figure 6.2, resembles a DAW, typically used in people's music making. IGME users can create and manipulate MIDI using generative AI systems (Hunt, 2021; Hunt et al., 2020). IGME helped people to ideate; however, only non-deep learning approaches were investigated (ibid). Calliope (Tchemeube et al., 2022) also resembles a DAW, providing an interface for uploading, editing and generating MIDI material from the Multi-Track Music Machine transformer model (Ens & Pasquier, 2020). A plugin with the same transformer model was also developed and tested within an existing DAW (Tchemeube et al., 2023); the user study showed that AIGC supported usability and people's feelings of ownership.

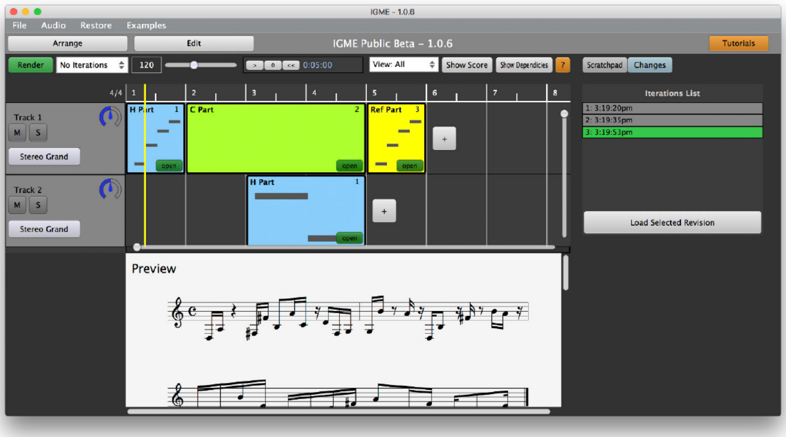
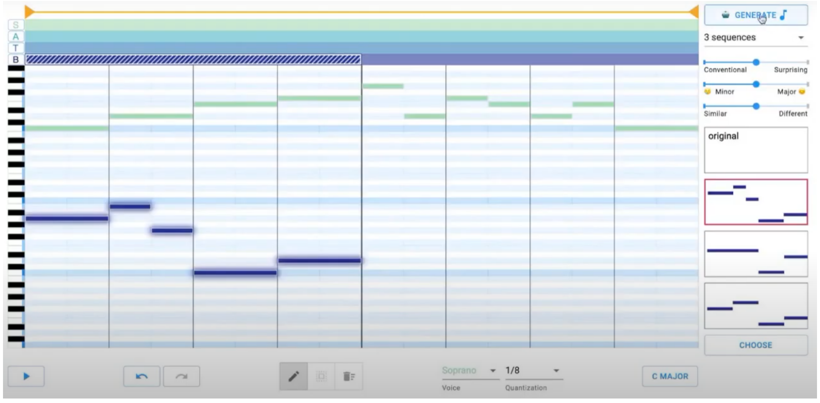


Figure 6.2: Interactive Generative Music Environment. Image from Hunt et al. (2020), CC BY 4.0.

[5] https://www.aisongcontest.com/

Controlled HCI studies have also investigated human-AI music composition using Cococo. Cococo is a web editor where users can write melodies and add an AI generated harmony part (Louie et al., 2020). The harmony can be steered towards conventional versus surprising outputs, and major versus minor outputs, using sliders (see Figure 6.3). Louie et al. (2020) found that the semantic sliders helped users feel ownership over the AI output and to express their musical intent. Louie et al. (2022) later showed that the expressive quality of their AI model also supported users' feelings of ownership and their musical intent. Ownership and agency are also raised as important aspects of composing music with AI in live coding contexts (Wilson et al., 2023; Xambó, 2022). Furthermore, pairs composing music with Cococo found the AIGC helped them to be playful as their critiques were expressed towards the AI and not each other (Suh et al., 2021).



3 sequences

Soprano
Voice

1/8
Quantization

Figure 6.3: Cococo's interface, where melodies can be augmented with AI generated harmonies. The sliders (right) can be used to steer the AI generated music output. Image from Louie et al. (2020), CC BY 4.0.

Several studies on AI-based music making have used FolkRNN, a neural network that generates folk music. Sturm (2022) showed that curating outputs from FolkRNN helped them express musical ideas they could not otherwise formulate. Sturm et al. (2019) found that FolkRNN shaped their music making as they negotiated compromises with the AI model. For example, whilst FolkRNN's temperature parameter was limited in how it could steer FolkRNN's output (all outputs were in the folk genre), it helped to generate surprising outputs at the edges of its training data. A notable output led one of the paper authors, Ben-Tal, to create a canonic piece, despite this being different to their typical composing style (Sturm et al., 2019).

Ben-Tal et al. (2021) also investigated how FolkRNN was used serendipitously when hosted on an online webpage. Users tended to repeatedly listen to FolkRNN outputs and select AIGC for their composition. Uncharacteristic ideas outside of the folk genre were later corrected by FolkRNN users for use in their compositions (Ben-Tal et al., 2021). This is similar to how Loth et al. (2023) used AIGC in the genre of progressive metal, modifying AI outputs which were unplayable or unnatural for the guitar. It is also similar to how musicians leverage imperfections or ambiguity as an aesthetic choice (Dannemann et al., 2023; Hamilton, 2020).

Gioti et al. (2022) reflected on the material and mediating properties of using AI in their music making practices. They described the need to modify their data collection process due to an AI's limitations. They also needed to modify several parameters of their AI systems to control the mapping between their controls and AI generated output. For example, they set up ways to react in real-time to the AI generation, such as by using faders on a mixing console.

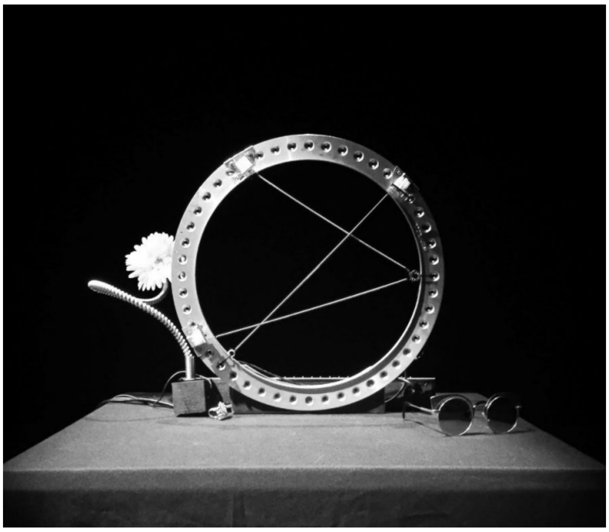


Figure 6.4: Laetitia Sonami's Spring Spyre. The metal springs stretched within the circular frame are mapped to audio output using interactive machine learning (Fails & Olsen, 2003). Image from Fiebrink and Sonami (2020), CC BY 4.0.

The rapid (re-)training of machine learning models to map gestures to musical outputs is central to the Interactive Machine Learning approach (IML; Fails and Olsen (2003)): training a machine learning model on a small amount of data based on judgments of how it performs in use. With IML, musicians often balance the accuracy of their mappings with the ability for mappings to create unexpected yet interesting outputs (Fiebrink et al., 2012). For example, Fiebrink and Sonami (2020) describe modifying training data to map metal springs (shown in Figure 6.4) to audio. Training with wide changes in the spring led to unpredictable outputs. Other systems use the ambiguity of the mapping between gestures and AI output. For example, Murray-Browne and Tigas (2021) mapped dancers' movement to areas of a latent space to generate music (see Section 6.2).

6.4 Summary of Literature Critiques

This chapter shows that there is little to no research on AI-based CSTs which directly address reflection, aside from a notable exception (Kreminski & Mateas, 2021). No studies have systematically evaluated reflection in the AI-based music composition domain. The section concludes the chapter with a summary of the literature critiques.

Music composition processes vary by genre (Biasutti, 2012; McAdams, 2004; McLean & Wiggins, 2010) and by the tools composers use (Sloboda, 1985; Webster, 2002; Younker, 2000). Although models of the composition process define broad stages, none indicate the types of reflection that occur within them. For example, preparation stages appear to include moments of reflection-for-action (Candy, 2019), yet this has not been systematically evaluated. This identifies a need to identify which types of reflection occur at different stages of the music composition process.

The use of AIGC further complicates the need to understand reflection across the stages of music composition. AI music systems have unique affordances that shape which types of reflection occur. For instance, rule-based systems and Markov chains (Carnovalini & Rodà, 2020) are interpretable, affording reflection on an AI's inner workings. However, more sophisticated deep learning systems (Briot et al., 2017) are challenging to understand, limiting reflection to their outputs rather than their inner workings. VAEs present new opportunities for reflection by mapping user gestures to compressed latent representations of musical data, such as movement (Murray-Browne & Tigas, 2021) or sliders (Louie et al., 2020). This ambiguity could present opportunities for reflection (Gaver et al., 2003), but further research is required to demonstrate this.

Existing HCI studies of the use of AIGC in creative practice, including for music, tend to focus on broad perceptions of AI, rather than specific types of reflection in CST interaction. There is emphasis on issues such as how much agency an AI has over a person's work (Amershi et al., 2019; Lewis, 2023; Louie et al., 2020; Wilson et al., 2021; Xambó, 2022), or how much people compromise on their existing creative process when using AIGC (Ben-Tal et al., 2021; Fiebrink & Sonami, 2020; Gioti et al., 2022; Sturm et al., 2019). Further work is needed to understand how these concerns characterise reflection during composition.

Some AI tools outside the CST field have shown potential for reflection support. For example, AI tools have prompted users to motivate their reflection during interaction (Hubbard et al., 2021; Li et al., 2023; Reicherts et al., 2022; Wagener et al., 2023). However, AI that interrupts users may disrupt flow states (Csíkszentmihályi, 1990). This demonstrates a need for investigations on the distinction between reflection and engagement for AI-based CST contexts.

Overall, few AI-based CSTs have directly assessed reflection, and none have systematically evaluated reflection in the context of music composition. There is an opportunity to directly investigate how people reflect when using AIGC in music composition. The following chapter thus presents a mixed-methods user study, including use of RiCEv2, to evaluate reflection across different AI-based CSTs.

Chapter 7

Artist-Researchers' First-person Reflections in AI-based Music Composition

The previous chapter showed that there are few studies on how people reflect using AI Generated Content (AIGC) in Creativity Support Tools (CSTs), and none systematically evaluating reflection in the music composition context. This chapter thus investigates the plurality of ways in which people reflect in AI-based music composition. It presents a collection of six first-person reflective accounts from artist-researchers on their experience of composing a piece of music, each using a different AI tool. The subjective accounts are combined with interviews and subjected to a Thematic Analysis (Braun & Clarke, 2006) to identify common reflection patterns. RiCEv2 questionnaire measures collected throughout each artist-researcher's composition process are also collected and analysed in the following chapter. This leads to the first systematic evaluation of reflection in AI-based music composition.

This chapter is structured as follows. The novel study method is introduced in Section 7.1 followed by its findings in Section 7.2. A discussion of the findings within the context of related literature closes the chapter (Section 8.3).

7.1 Method

Qualitative and quantitative data were collected to showcase individual insights and identify commonalities in AI-based music composition. The study was inspired by ethnographic approaches (Benford et al., 2013; Chamberlain et al., 2012; Millen, 2000) to evaluate a range of composition practices in their usual locations of happening (see Section 2.3.1.3), such as in the home (Benford et al., 2013; Wilson et al., 2023). The methodological novelty is to purposefully ask the composers to pause and reflect back on their music making, documenting their thoughts using a *reflection board* (see Section 7.1.4.2).

The study was approved by the Queen Mary University of London ethics committee. Participants provided written consent and were reimbursed with a £100 (GBP) voucher, following pricing in the range of the UK Musicians' Union's rates for commissioning 1 minute of music for media[1]. Each participant was acknowledged as a co-author for a publication about this work.

7.1.1 Participants

To recruit participants, e-mails were sent to research groups in the UK with interests in music and AI. This group was chosen to target participants with the technical skills to use state-of-the-art AIGC in a music practice and academic writing skills for the first-person accounts. The criteria for participation were to be a PhD student, have developed a way of integrating AI in music making, and be eighteen or older. The participants are thus composers and artist-researchers (Sturdee et al., 2021). Their perspective is unique in that they think about AI music in their everyday work life.

Seven composers were recruited in total. Two collaborated on a single composition as a band. The choice was made to include the band to be able to give insights into an example of collaborative music making. The participants' characteristics are shown in Table 7.1, drawn from a questionnaire which included the MSI (Müllensiefen et al., 2014) to assess musical expertise and SRIS (Grant et al., 2002) to assess their capacity for self-reflection. For the MSI scores (Mean = 110.1, Med = 112, SD=8.6), all participants are above the UK average (86). For the SRIS scores (Mean = 5.1, Med = 5.3, SD = 0.4), 6 out of 7 participants scored higher than the CSTs users from Chapter 3 (4.5); P1 is close to this average, scoring 4.3. Further details are introduced throughout Section 7.2.1 for context.

[1] https://musiciansunion.org.uk/working-performing/composing-and-songwriting/commissioned-work/media-commissions

Table 7.1: **Overview of the artist-researchers' characteristics.**

ID	Age	Gender	Self-described Music Experience	MSI Score	SRIS Score
P1	26	Male	Performed in electronic and contemporary music ensembles for 10 years. Masters in Sonic Arts. Plays guitar and drums. Previous experience writing contemporary and minimalist music for chamber groups, jazz, indie and popular acts.	114	4.3
P2	25	Female	Undergraduate degree in Creative Music Technology. Media composer writing music for published video games, short movies and media companies. Also worked writing music for dance performances.	115	5.3
P3	26	Male	Guitarist for 15 years. 6-7 years music composition and production experience. Has released 5 original albums and produced/mixed music. Played in rock bands on guitar, bass, drums and vocals. **Note:** in a band with P4.	112	5.0
P4	32	Male	20+ years experience composing music, from classical guitar pieces to progressive metal. Experience as a solo classical guitarist and in 5 people ensembles (drums, two guitars, bass, keyboards). **Note:** in a band with P3.	109	5.2
P5	31	Female	Writing music for 15 years using conventional instruments e.g. guitar and piano. 5+ years experience as a live coder, making experimental electronic music, actively gigging.	106	5.5
P6	29	Male	Classically trained composer, writing both as a traditional composer and working with various small ensembles. Also a performer/improviser. Actively gigs. Writes experimental and computer music, and contemporary classical.	121	5.3
P7	29	Male	BA (Hons) in Creative Music Production; MSc in Sound and Music for Interactive Games. Specialised in composing for games. IMDb credit for a feature length horror film. 15+ years experience as a performer in death metal bands.	94	5.3

7.1.2 AI Tools

Each music composition in this study was made with a different AI tool. The participants self-selected the AI tool and decided how to integrate it into their workflow. A deliberate choice was made to study AI tools from the research community instead of commercial work to consider the emerging state-of-the-art (no participants chose to use commercial AI tools). The AI tools were considered sufficient for use if they had been used in music making previously (as evidenced by showing previous examples of their own or others' compositions with the AI tool) or published at an academic conference. All participants selected AI tools they had used at least once before to make music.

Three AI model architectures were present in the selected tools. As described in Section 6.2, these were Markov Chains, Transformers, and Variational Auto-Encoders (VAEs). The AI tools selected are summarised in Table 7.2, with screenshots in Figure 7.1. For context, further details on each tool are given before each first-person account in Section 7.2.1.

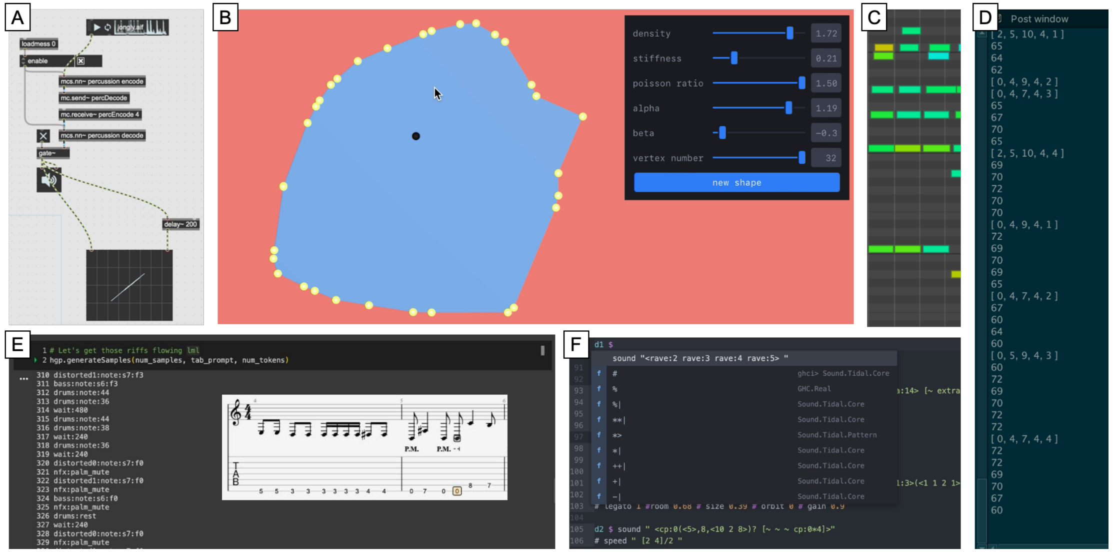
7.1.3 Procedure

Participants were asked to freely write a music composition with a minimum length of one minute in their chosen genre, using their chosen AI. They were asked to complete four sets of one-hour composition sessions, pausing to reflect on their composition after every hour (in pilot tests, four hours was sufficient for a full composition cycle from ideation to completion for one minute of composed music). Coincidentally, all participants requested to complete the sessions within one day to balance with their other time commitments.

The study was completed remotely to allow participants to be located in typical environments for their music making (Benford et al., 2013; Wilson et al., 2023). The choice to create moments for the composers to pause and reflect every hour, instead of composers self-selecting moments to reflect, was to ensure that sufficient data was captured on people's reflection, whilst being mindful of time constraints. This contrasts methods for CST studies on qualities such as feelings of flow states (Csíkszentmihályi, 1990), where interruptions would pose a confounding variable. Here, interruptions are

Table 7.2: Summary of the artist-researchers' chosen AI tools. Labelling (a) to (f) refers to the image labelling in Figure 7.1.

AI	Architecture	Input	Output	Integration
RAVE (P1)	VAE	Audio	Audio with modified timbre	Plugin for Max
Neural Resonator (P6)	Neural Network	Audio excitation & UI	Audio of a synthesised drum	Plugin for music software
CFEP (P7)	Transformer	Text (MIDI)	Humanized MIDI as Text	Manual import MIDI
Mark of Markov (P2)	Markov Chain	Manual parameters in code	MIDI Notes and chords	Records to music software
ProgGP (P3 & P4)	Transformer	Text (Guitar Tab)	Text (Guitar Tab)	Manual import MIDI
Tidal-Fuzz (P5)	Markov Chain	Text (Music Code)	Text (Music Code)	Manual import MIDI



density 1.72
stiffness 0.21
poisson ratio 1.50
alpha 1.19
beta -0.3
vertex number 32

```
[ 2, 5, 10, 4, 1 ]
65
64
62
[ 0, 4, 9, 4, 2 ]
[ 0, 4, 7, 4, 3 ]
65
67
70
65
[ 2, 5, 10, 4, 4 ]
69
70
72
70
70
[ 0, 4, 9, 4, 1 ]
72
69
70
69
65
[ 0, 4, 7, 4, 2 ]
67
60
64
60
[ 0, 5, 9, 4, 3 ]
60
72
69
70
72
72
[ 0, 4, 7, 4, 4 ]
72
72
69
70
67
60
```

```
# Let's get those riffs flowing lml
hgp.generateSamples(num_samples, tab_prompt, num_tokens)
```

```
310 distorted1:note:s7:f3
311 bass:note:s6:f3
312 drums:note:44
313 drums:note:36
314 wait:480
315 drums:note:44
316 drums:note:38
317 wait:240
318 drums:note:36
319 wait:240
320 distorted0:note:s7:f0
321 nfx:palm_mute
322 distorted1:note:s7:f0
323 nfx:palm_mute
324 bass:note:s6:f0
325 nfx:palm_mute
326 drums:rest
327 wait:240
328 distorted0:note:s7:f0
329 nfx:palm_mute
```

```
d1 $
sound "<rave:2 rave:3 rave:4 rave:5> "
f  #     ghci> Sound.Tidal.Core
f  %     GHC.Real
f  %|    Sound.Tidal.Core
f  **|   Sound.Tidal.Core
f  *>    Sound.Tidal.Pattern
f  *|    Sound.Tidal.Core
f  ++|   Sound.Tidal.Core
f  +|    Sound.Tidal.Core
f  -|    Sound.Tidal.Core
a:14> [~ extra
1:3>(<1 1 2 1>
# legato 1 #room 0.68 # size 0.39 # orbit 0 # gain 0.9

d2 $ sound " <cp:0(<5>,8,<10 2 8>)? [~ ~ ~ cp:0*4]>"
# speed " [2 4]/2 "
```

Figure 7.1: AI tools integrated into the composers' practices. (A) Real-time Audio Variational auto-Encoder (RAVE) (Caillon & Esling, 2021); (B) Neural resonator (Diaz, 2024). (C) Cue-Free Express + Pedal (CFEP) (Worrall et al., 2022); (D) Mark of Markov (MoM); (E) ProgGp (Loth et al., 2023); (F) Tidal-Fuzz (Wilson et al., 2021).

used to scaffold reflection and moments of engagement are not a focus of the study.

The steps of the procedure are as follows:

1. **Pre-task Questionnaire:** Participant completes the pre-test questionnaire (10 minutes, see Section 7.1.1).
2. **Task:** Participant composes with their chosen AI whilst recording their computer screen for 60 minutes.
3. **Pause:** The researcher notifies the participant that the 1 hour session is finished.
4. **Questionnaire:** The participant completes the RiCEv2 questionnaire for the recent session (see Section 7.1.4.1).
5. **Reflection Board:** The participant completes a reflection board for 30 minutes (see Section 7.1.4.2).
6. **Interview:** The participant and researcher meet for a semi-structured interview online for 10 minutes (see Section 7.1.4.3).
7. **Repeat:** Steps 2 through 6 are repeated until 4 sessions are completed.

7.1.4 Data Collection

A mixed methods approach was used as per the thesis methodological approach (see Section 1.4). Several data sources were collected after each one hour music making session, including: reflection questionnaires, reflection boards and interview data. A first-person account from composers was collected after all sessions were completed. The rationale for this was to interpret the questionnaire measures within the richer context of participants' reflections through triangulation of the quantitative and qualitative data.

7.1.4.1 Reflection Questionnaire

To identify possible patterns in reflection throughout the composition processes, metrics were gathered from RiCEv2 (see Table 5.2). To recap, averages from these statements are calculated for: reflection-on-current-process, reflection-on-self, reflection-through-experimentation and a total RiCEv2 score. These are analysed in Chapter 8.

7.1.4.2 Reflection Boards

Participants were given a template for the online collaborative whiteboard Miro[2] (see Figure 7.2). The template posed questions at the top of a set of columns based on the three factors of RiCEv2 to prompt and organise the participants' thinking. Participants were instructed as follows:

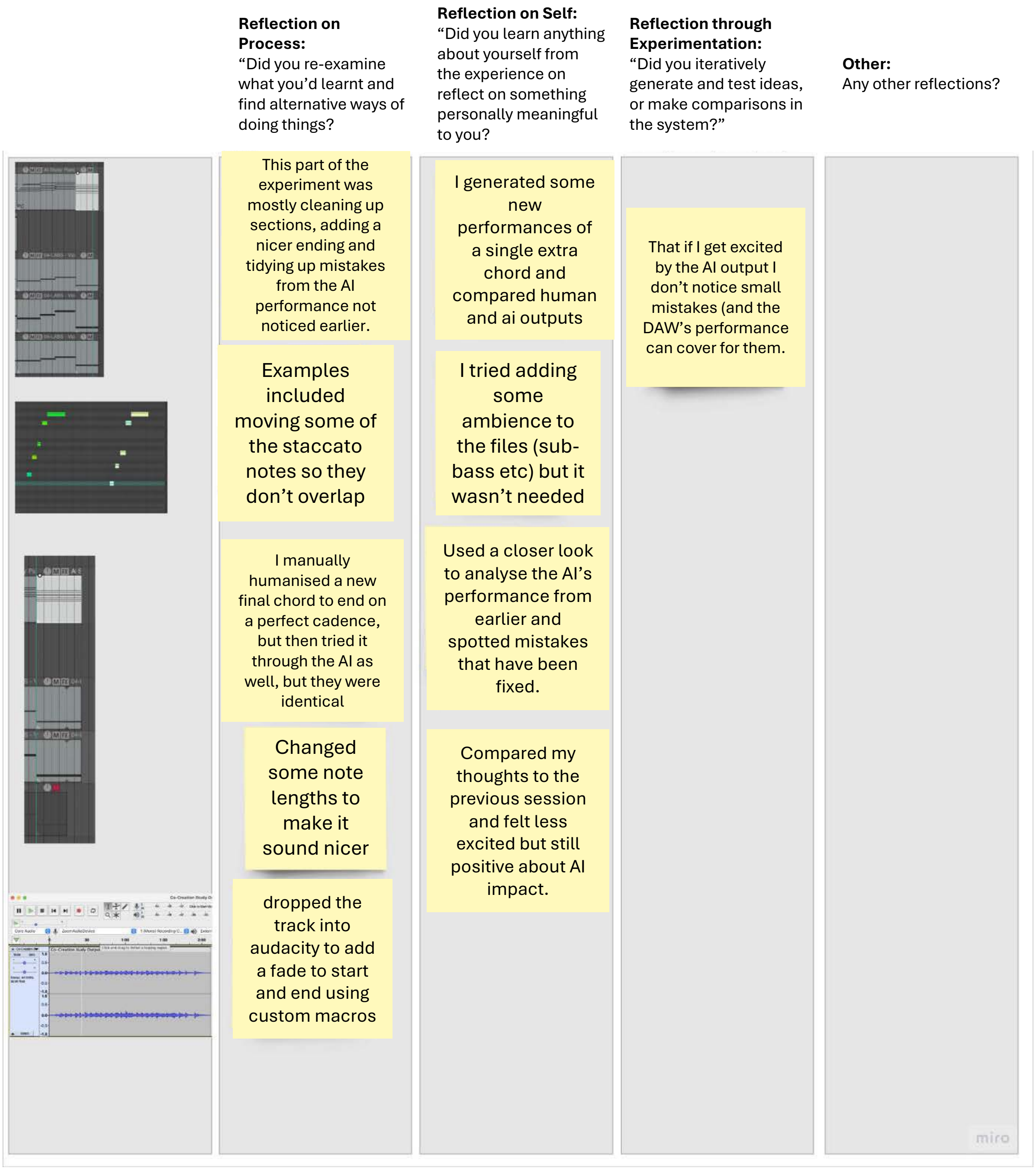
> "Add 6 to 10 screenshots from your video recording into Miro that best represent your creative process in the session. Organise them in chronological order in the leftmost column. Then, use the post-it note feature (press 'N') to document your reflections and thoughts during your composition process. Use the questions at the top of each column to prompt and organise your thinking. I'd expect to see near 10+ post-it notes. This should not take long, at most 30 minutes."

Participants were asked to document screenshots in chronological order (from top to bottom) in the leftmost column to show how the composition unfolds, as inspired by studies on the composition process (Collins, 2007; Folkestad et al., 1998). Screenshots were used as they offer insights into the participants' personal decisions at specific points in time (Gamboa et al., 2023). Then, using the post-it note feature, participants were asked to document their reflections and thoughts on their composition process, using the guiding questions at the top of each column. The reflection boards were used instead of other retrospective protocols (Candy et al., 2006) such as video-cued recall, so that the composers could be self-sufficient in their documentation and quickly refer to the data later for their first-person accounts.

7.1.4.3 Interview

A short interview was undertaken in which participants were asked to talk through what they did in the preceding hour, and then talk through the reflections in Miro. The approach was semi-structured to give the researcher opportunities to probe unexpected lines of discussion. The questions are

[2] https://miro.com/



Reflection on Process:
"Did you re-examine what you'd learnt and find alternative ways of doing things?

This part of the experiment was mostly cleaning up sections, adding a nicer ending and tidying up mistakes from the AI performance not noticed earlier.

Examples included moving some of the staccato notes so they don't overlap

I manually humanised a new final chord to end on a perfect cadence, but then tried it through the AI as well, but they were identical

Changed some note lengths to make it sound nicer

dropped the track into audacity to add a fade to start and end using custom macros

Reflection on Self:
"Did you learn anything about yourself from the experience on reflect on something personally meaningful to you?

I generated some new performances of a single extra chord and compared human and ai outputs

I tried adding some ambience to the files (sub-bass etc) but it wasn't needed

Used a closer look to analyse the AI's performance from earlier and spotted mistakes that have been fixed.

Compared my thoughts to the previous session and felt less excited but still positive about AI impact.

Reflection through Experimentation:
"Did you iteratively generate and test ideas, or make comparisons in the system?"

That if I get excited by the AI output I don't notice small mistakes (and the DAW's performance can cover for them.

Other:
Any other reflections?

Figure 7.2: An example reflection board from the last session completed by P7.

shown in Table 7.3. The aim was to elicit descriptive accounts of the participants' experience and to clarify the participants' reflections in Miro. This justifies that the questions used in the interview were the most open-ended, compared to the later studies in this thesis. The interviews also served as a contingency in case the participants could not complete their first-person accounts.

Table 7.3: Interview questions used alongside the reflection board process in the artist-researcher user study.

Composition Session Questions
- Talk me through your composition process and what you did in the session.
- Talk me through the reflections that you have written in Miro.
- Was there anything that stood out as very important or very surprising?
- If you had to summarise what you just did in your composition session in one word, what would it be?

7.1.4.4 First-person Account

After all sessions were completed, participants wrote an 800-1000 word **first-person account** with the following instructions:

> "Write an account of how you composed with your chosen AI and what you reflected upon, looking over your Miro boards. We only expect first-draft quality. Please include all the key points you would like to talk about that you think are important, using your own voice."

7.1.5 Data Analysis Method

The reflection boards, interview transcripts, and first drafts of the first-person accounts were collated into one document to consider the variation of data together (Millen, 2000). The transcription was conducted using automatic captioning, and then correcting errors manually (the author reads the transcript over several passes whilst listening to recordings). The transcripts are at the level of detail that allows researchers to understand the discussions from the text alone, for example, by annotating non-verbal communications in brackets. The transcripts are not to the level of detail for conversation analysis (Hepburn & Bolden, 2012), for example, including umms and errs.

An inductive thematic analysis (Braun & Clarke, 2006; Braun & Clarke, 2019) approach was performed on the collated data. Thematic analysis was selected because it produces a good level of descriptive detail whilst being manageable by an individual researcher (Bryan-Kinns et al., 2018). Specifically, the *reflexive* thematic analysis approach (Braun & Clarke, 2019) was followed: the author does not take a passive role in the analysis, instead generating themes by reading the data through their own experience and knowledge of underlying theories. Themes do not *emerge* from this analysis approach, but the researcher actively co-constructs the themes with the data (Braun & Clarke, 2019, 2021). An *inductive* reflexive thematic analysis is used instead of applying a pre-existing coding scheme to identify unanticipated findings, which are likely given the open-ended study task used.

The thematic analysis approach moves back and forth between the steps of: familiarisation (the researcher reads through the transcripts with audio over several passes, immersing themselves in the data, and jotting initial thoughts); coding (identifying and labelling key moments of interest); refining codes (revisiting the codes and identifying patterns across codes or potential deeper insights); themes (clustering the codes by their shared meaning and underpinning concepts); theme fit (revisiting and verifying codes against the theme definitions). The rigour and consistency of the thematic analysis were verified through regular team meetings and by comparing the codes against the thesis's research questions.

7.2 Findings

The following subsections report summaries of the first-person accounts and the thematic analysis findings. These are included for each participant to report their personal and nuanced perspective (Ellis et al., 2011; Fdili Alaoui, 2023). The full-length first-person accounts are in the appendix.

7.2.1 Summary of First-person Accounts

For each first-person account below, a reminder of each composer's expertise and details on their chosen AI tool are given. The first-person accounts are then presented, edited to best relate to the thesis research questions. Readers are invited to listen to each music composition online[3].

7.2.1.1 P1: Ash

Ash composes music with a glitch aesthetic, recording improvisations with interfaces they create using the visual programming language Max[4]. They chose the VAE model named RAVE (Caillon & Esling, 2021) (see Figure 7.1a). RAVE can take an audio clip as input and change its timbre. For example, a recording of a person singing can be transformed to sound like a trumpet following the same melody. RAVE can generate high-quality 48kHz audio signals and be used with a standard laptop's central processing unit (Caillon & Esling, 2021). RAVE can also be controlled by varying values of the latent space in its VAE architecture and feeding this into its decoder.

> "Typically, I like to get output as soon as possible, but I was surprised by how little I initially got from RAVE. The 8-dimensional input of RAVE and its non-deterministic output made me reevaluate the structure of my typical process. I considered ideas from John Croft (2007), such as what layer of abstraction (or the level of complexity) I wanted.
>
> Through various signal processing techniques, I ended up with a way to control both RAVE and a non-AI FM synth. This allowed me to negotiate between the AI and non-AI sounds, where you can decide which to dominate whilst improvising. The combination of predictable and unpredictable, semantic and black box, brings a similar level of expectation with pleasant surprise as I had experienced being in jazz ensembles. However, I still couldn't think of my composition in a deterministic way, like in FM synthesis where you have a good idea of what will happen when a parameter is changed (see Stria by Chowning (1977)). I can't control the model and know what it's doing, so I handed off control to the AI."

[3] https://codetta.codes/reflection-across-AI-music/

[4] https://cycling74.com/products/max

7.2.1.2 P2: Sara

Sara is a media composer with experience working for video game companies, mainly writing orchestral music. Their chosen AI *Mark of Markov* (MoM; see Figure 7.1d) uses Markov Chains to output notes and chords that switch between modes (scales offering different musical moods) based on various probabilities[5]. Each chord output is a bar in length. On compiling MoM, its output is synthesised in real-time and recorded as MIDI.

> "The initial material generated by MoM was boring – too quantised and not human at all. Because MoM spits out MIDI based on its previous music, I couldn't copy and paste parts from the melody and stitch them together, because there is a chance the chords could be in a different key.
>
> I felt really bad changing the stuff MoM created – I wanted to use all of it so it did not go to waste. I thought that if I kept changing the system output, was I really using it to its full potential? Was I just taking over?
>
> Whilst composing, it was interesting that I kept making comparisons to a composition I previously wrote using MoM, which I was really proud of. I also would compare myself to people such as John Williams[6], and think, "well if I am going for a similar style to his, I cannot even get close to the quality of his compositions". This can get very demoralising and add a lot of pressure. I found taking small chunks of the output and trying to make them work together helped to take off the pressure."

7.2.1.3 P3 and P4: HEL9000

Jack (P3) and Pedro (P4) create progressive metal music using AI, as the band HEL9000[7]. They chose ProgGP (Loth et al., 2023), a transformer model (Dai et al., 2019) trained on the DadaGP dataset (Sarmento et al., 2021) – a dataset of 26,000 rock and metal guitar tablatures[8] – and fine-tuned further on a set of progressive metal guitar tablatures. The notation

[5] MoM is described at: https://saracardinalemusic.com/project/mark-of-markov/

[6] https://www.imdb.com/name/nm0002354/

[7] https://twitter.com/HEL9000ismetal

[8] Guitar tablatures are a music notation system designed specifically for guitarists.

software Guitar Pro[9] is used by HEL9000 to write guitar tablatures, which are converted to text and fed as a prompt to ProgGP in a Google Colab notebook[10] to generate continuations of rock and metal songs (Sarmento et al., 2023) (see Figure 7.1e). Notably, outputs contain not only guitar sections, but also bass and drums alongside the guitar, and are converted to MIDI to be added to music software for editing.

> “The interaction with ProgGP was mostly dictated by an initial need for isolated riffs, or musical ideas, that could be put together to form a full song. The process started with Jack experimenting on guitar to compose a riff.
>
> After Pedro notated the initial riff into tabulature manually, we input the riff to our AI. We divided our workflow: Pedro took care of filtering continuations and feeding them back into the model to get variations; Jack started recording the initial riff on guitar to the computer, and adding drums and bass digitally. After Pedro filtered ideas, we both listened to the AI outputs together and curated a few riffs we felt could be put together coherently.
>
> We then focused on recording these ideas. To enrich the song, we added extra layers using samples or new lead guitar parts. One particular AI output had a distinctive drum beat generated alongside the guitar riff, which prompted us to explore samples that we wouldn’t usually use for [the band’s] music. Another section made us reflect on The Ocean’s[11] aesthetics, prompting us to include a marimba and glockenspiel over a lead guitar part. Inspired by Periphery[12], we added a piano mimicking the melodic line of the guitar.”

7.2.1.4 P5: Lizzie

Lizzie creates experimental electronic dance music as a live coder, where code is executed in real-time to produce sound and music. They use the domain-specific programming language Tidal Cycles (McLean & Wiggins,

[9] https://www.guitar-pro.com/

[10] https://colab.research.google.com/

[11] https://www.theoceancollective.com/

[12] https://periphery.net/

2010), an extension of the functional programming language Haskell. Their chosen AI, Tidal-Fuzz, is a Markovian agent that outputs code sequences by randomly walking through and choosing Tidal Cycles functions that form musical patterns (Wilson et al., 2021). These are integrated into the user interface as suggestions to add to the music code cf. GitHub Co-Pilot. The code produced by Tidal-Fuzz uses Haskell's strict type system to ensure code is syntactically correct and executable.

> "Where patterns were solely created by the human, some reflection came through errors made. For example, at one point, I was looking for a specific sample and typed the wrong number, which prompted me to explore a sample that I might've not considered.
>
> With patterns solely created by Tidal-Fuzz, reflection materialised in a few separate ways. Firstly, the agent's patterns were evaluated against my aesthetic preferences. A lot of reflection occurred around evaluating whether these matched *my* aesthetic preferences. I had an internal aesthetic function in mind to try to express a specific affective state, which I tested against the AI's aesthetic function to see how they matched.
>
> In understanding the affective states driving my internal aesthetic evaluation function, through considered, deep listening (see Oliveros (2005)), I also was forming understandings of myself in relation to the music."

7.2.1.5 P6: Lewis

Lewis is a composer and performer, including in the band Julia Set[13]. They typically create experimental computer music and contemporary classical. They chose the neural resonator plugin (Diaz, 2024) (see Figure 7.1b), which uses neural networks to predict coefficients for a resonant filter bank (Diaz et al., 2022). An audio or MIDI excitation is input to the plugin and used to trigger feedback, propagating throughout the filter bank to synthesise different drum sounds. Moving parameters on the plugin's interface changes the shape of the drum (that is, it modifies the filter bank coefficients). A button is also provided which randomly chooses a set of parameter values.

> "As much as I was familiar with the Neural Resonator already,

[13] https://juliaset.bandcamp.com/

> I was not able to clearly audiate (meaning to imagine sounds mentally) its product – a familiar challenge I've tackled in my compositional work. This brings forth differences between my mentality as a composer (how do my actions affect my future self and what is my creative idea) and as an improviser (how do my actions affect my *present* self and what is the *performed* idea/instinct). I spent the second composition session generating material by improvising with the AI, using instinct. This enabled me to generate a large amount of material, creating many threads of ideas from which to develop a composition.
>
> In the third composition session, my mentality shifted away from the instinctual and towards the considered. My creative decisions were no longer influenced by the AI, but were instead imposing themselves onto the material it had just generated.
>
> My creative ideas [arose most easily] through listening/interpreting than conceiving/enacting. I could form an expression of the self through my previous instinctual responses and explorations."

7.2.1.6 P7: Kyle

Kyle is a media composer specialising in game audio. They chose the AI, CFEP (Worrall et al., 2022), which transforms MIDI recordings to sound more human and expressive, based only on the musical features of pitch and note timing; this supports expressiveness when richer musical data is not available (Worrall et al., 2022). It combines transformer models trained on piano datasets to predict the velocity, timing and tempo for input MIDI files, outputting more human sounding adjustments of the input music (see Figure 7.1c).

> "I couldn't use CFEP without having written initial material. It was interesting how the AI coloured my initial choices. I call this a butterfly effect, where CFEP's design had unexpected knock-on impacts on my creative workflow. The first butterfly effect was in choosing piano – although common to my chosen genre, I also chose the instrument knowing that CFEP is trained on a piano dataset, so it would perform well on this type of data.
>
> In session two, I began to experience the AI as a pseudo-

> co-producer, in the sense that the inclusion of it in the project influenced decisions that you make creatively. For example, I added staccato piano notes and drums. However, I eventually disregarded these ideas because, in addition the vibe/feeling of the music not being correct, I knew CFEP would ultimately not work well on drums.
>
> Surprisingly, I thought that the AI output was good enough that I felt moved. I really did not expect to be moved by the piece and I do not know why I found this quite moving."

7.2.2 Thematic Analysis Findings

Six themes were generated from the thematic analysis of the participants' first-person accounts and interviews, described below.

7.2.2.1 Theme 1: Reflection on Past Instincts

P1 and P6 curated AIGC by reflecting "in the moment" (P6) and using their instincts. They created environments where they could listen to AIGC in real-time – in a way that was more "improvisatory" (P6) and "instinctual" (P6), to "try a bunch of stuff" (P1). P6 found this "felt quite familiar" to how they would improvise in their music practice. P1 said the process reminded them of their time playing in jazz ensembles.

Furthermore, P6 described this process as deliberate: they split their process into choosing material based on their instinct in Session 2 and then reflected on their *past decisions* when organising this material in Session 3. Indeed, noted P6 reflecting on their *future self* in their first-person account.

P5 also describes reflection-on-self when curating material by listening in real-time and live coding. They said they were "forming understandings of [themselves] in relation to the music" (P5), and reflected on how AIGC matched their aesthetic: "the things that [the AI] was producing weren't necessarily in my aesthetic, so then it was a case of refining what it was that I actually wanted" (P5).

7.2.2.2 Theme 2: Reflection on Direction and Surprises

All the composers reflected on the direction in which to take their music. For example, P5 found the AI "pushes me in different directions or gets me thinking about doing things in a different way that I haven't thought about myself". The direction for a piece was also considered through reflection-on-surprises, both from the AI and other aspects of composers' workflows. For example, P1 found their AI "really worked as a surprise prompt", helping them to continue "taking risks and experimenting". P5 tested a cowbell sample they usually would not use in their practice, noting that "this surprise moment was[...] crucial for building [their music]".

7.2.2.3 Theme 3: Reflection for AI

P1 and P7 reflected on how their current actions would integrate with their chosen AI tool. P7 described their AI as a producer – musical material fed to the AI would either work or not work. P1 explained that, similar to when you compose for performers and shape your composition to what people can physically play on their instruments, you shape your composition to the AI and its affordances – "you have to take into consideration things like what people can physically play[...] so they kind of shape your composition because of the limitations[...] I think it's quite similar [with AI]" (P1). P7 went so far as to describe this as a butterfly effect, where their compositional decisions were limited to those that would work well from the start with their chosen AI.

7.2.2.4 Theme 4: Reflection on Feelings

P2 and P7 reflected on their feelings of using the AI in their practice. P2 felt bad about changing outputs from their AI system. They described self-awareness and feelings of impostor syndrome, such as when comparing themselves against the famous composer John Williams in their first-person account. A different emotional response was from P7, where their chosen AI, CFEP, transformed their music to sound more humanistic, and they were "surprisingly moved" that the AI could play their music in a way they felt they could not.

7.2.2.5 Theme 5: Reflection on Influences

The first-person accounts identified several inspirations that the composers reflected on to inform their creative practice. There were references to literature from musicians and philosophers such as Croft (2007), Chowning (1977), Oliveros (2005) and Periphery (see P3 and P4's first-person account).

It was possible to trace the musicians' creative influences to ideas implemented into their practice. For example, P1's negotiation between AI and non-AI mirrors Croft's (2007) philosophising on levels of control to afford in musical improvisations. P1 also cited Chowning (1977) to give an analogy to how the unpredictability of AI outputs was different to music making without AI: "I[...] couldn't think of my composition in a deterministic way, like in FM synthesis".

P3 and P4's choice of instruments was based on the influence of different bands, such as piano inspired by Periphery, and marimba and glockenspiel inspired by the Ocean. P5's reference to Oliveros's (2005) concept of deep listening gave insights into their understanding of how they reflect when evaluating their music, such as by reflecting on their "affective state" and "aesthetics". This relates to their reflection-on-self: "I[...] was forming understandings of myself" (P5).

7.2.2.6 Theme 6: Reflection on Technical Challenges

The composers reflected on challenges in integrating their chosen AI into their workflow. P1, P3, P4, and P7 mentioned the need to format data to move between their instruments and the AI inputs and outputs. For example, P3 and P4 needed to notate ideas they had developed on their guitars by "tap[ping] it out manually" (P3) and then converting the notation into a symbolic format, before being able to feed their data to their AI.

P1, P2 and P6 used templates with the AI already within their composition software to avoid complex setup. For example, P2 "opened the old projects where I knew that the output of the code was strictly going straight to Logic to record, because otherwise I would have to change inputs and outputs, and I don't remember how to do it."

P1 thought their AI was unimpressive to start with and spent lots of time tinkering before getting interesting results – in turn worrying about "sunk costs" where you waste "loads of time into making the system work." P2, P6 and P7 tweaked their code whilst music making – for example, P7 had "a little section in the middle where I edited some code [...] to change the variable name to the file name".

7.3 Discussion

Table 7.4: Summary of Chapter 7's main findings.

Main Finding	Location
Composers used AI in ways that were familiar to them to support their reflection. For example, composers would select AIGC whilst listening in real-time, to be closer to their improvisation practice.	§7.2.2.1
Reflection is characterised in AI-based music composition as including moments of reflecting on your future self, when listening to AIGC being generated in real-time.	§7.2.2.1
Reflection is characterised in AI-based music composition as including moments of self-reflection when organising AIGC, such as whether the AIGC fits an artist's aesthetic.	§7.2.2.1
Reflection is characterised in AI-based music composition as including moments of reflecting on the limitations of an AI and tweaking musical material to fit. For example, deciding to input piano melodies instead of drums for an AI trained on piano data.	§7.2.2.3

This chapter shows a collection of first-person accounts demonstrating the plurality in people's reflection across different AI-based music composition practices. Six music compositions were written by artist-researchers integrating a unique AI tool of their choice into their typical music making practice. Common themes were shown through a thematic analysis of these accounts and interviews conducted throughout the composition process. The main findings are summarised in Table 7.4. The findings are discussed in context with related literature below, followed by a discussion of the study's limitations.

The first-person accounts enabled investigation into various individualistic reflections on AI-based music composition. For example, P7's chosen AI, CFEP (Worrall et al., 2022), uniquely gave an example of an AI which required musical material to be written beforehand.

P3 and P4 offer unique insights as an example of collaborative practice. It was observed that they split tasks to effectively navigate the co-creative AI process (Huang et al., 2020; Muller et al., 2020). P4 initially generated ideas, whilst P3 prepared the music software cf. reflection-for-action (Candy, 2019). It is notable that they completed the curating of AIGC together. This emphasises the importance of selecting AI outputs, as this had significance to both band members. It also demonstrates how P3 and P4's composition process was adapted to be more familiar to them, mimicking how rock bands make music without AI by jamming (Biasutti, 2012). This also mirrors how Yurman and Reddy (2022) assigned meaning to AIGC in their art making context.

P2 was unique because their AI tool, MoM, did not require input and only output music to their software. This informed their comments that composing with smaller chunks of AIGC was easier. For P2, editing the recording to curate different ideas was necessary to identify interesting motifs. This approach was also more familiar to them, as common in film music making practices (McAdams, 2004).

Whilst P2 does not mention challenges presented by tools that require input, others found they had to adapt their practice to input material appropriate for their chosen AI in Theme 3 (Reflection for AI). For example, P7 avoided writing material for drums as CFEP was trained on a dataset of piano music. They thus thought CFEP would not perform well on melodic data. This demonstrates how the AI shaped their musical practice, corroborating how Gioti et al. (2022) and Sturm et al. (2019) negotiated with AI in their music making practices.

A notable finding from Theme 4 (Reflection on Feelings) related to two participants' unique emotional responses to the AI. P2 suggested that their AI helped them overcome impostor syndrome by providing material to extend. A different emotional response was from P7, who notes that their AI played music at a higher standard than themselves, helping them realise their music beyond their abilities. This corroborates Sturm (2022), who found their AI helped them to express ideas they felt they could not yet realise.

7.3.1 Limitations

The first-person accounts offer insights into AI-based music composition, but include many variations – from participant background, to the tools used, to the compositional techniques, and different genres. The findings *do not* generalise. However, they successfully capture qualities of a plurality of the artist-researchers' (Sturdee et al., 2021) real-world practice – specifically, for the few UK PhD researchers recruited. The approach of first-person accounts also suits making practices where a heterogeneity of different tools is the norm. For example, there is a range of tools within the live coding community (Aaron & Blackwell, 2013; McLean & Wiggins, 2010; Wilson et al., 2021). Future work can refine the findings further by focusing on different groupings of tools, for example, examining only the timbre-focused tools used, such as RAVE (Caillon & Esling, 2021) and neural resonator (Diaz, 2024).

The thematic analysis grouped the data to find commonalities, thus missing nuances on the differences exposed in the first-person accounts. Future work could adopt analysis techniques which embrace individualistic differences in perspectives. For example, a diffractive analysis approach (Morrison & McPherson, 2024; Nordmoen & McPherson, 2022; Rajcic et al., 2024; Robson et al., 2024) would allow for a deeper investigation of participants' musical practices, giving space to identify the entanglement between their similarities and differences in background, motivation and interaction styles. On the other hand, thematic analysis is a repeatable, systematic approach that supports the reproducibility of the findings.

Without conducting the interviews, participants could have completed the study at any time without the researcher needing to be present. In this case, participants could choose which moments to pause and reflect on whilst composing. This would have been more conducive to investigating aspects of creative user experiences, where interruptions pose a confounding variable such as flow states (Csíkszentmihályi, 1990). On the other hand, the structured approach to the music making activity used in this chapter meant that the collected data captured unique perspectives from artist-researchers with limited time for creative activities (Sturdee et al., 2021). The interviews also acted as a useful contingency in case participants could not complete their first-person accounts.

The study's method contrasts other methods such as diary studies (Botella et al., 2019; Dalsgaard & Halskov, 2012) or autoethnography (Lewis, 2023; Lucero, 2018; Noel-Hirst & Bryan-Kinns, 2023; Spiel, 2021). With these approaches, more commitment is typically required from participants. Diary studies also tend only to capture immediate thoughts. The reflection board method required participants to synthesise their thinking into first-person accounts retrospectively. This gave space for them to articulate their findings (Edmonds, 2022).

Future research could refine the reflection board technique by applying more time-deepening strategies (Millen, 2000). Nonetheless, collecting and comparing a range of first-person perspectives was helpful. It successfully captured multiple personal insights in a systematic way. This complements research using more common HCI methods (Nicholas et al., 2022).

7.4 Conclusion

This chapter showed how composers reflect across a range of AI music tools and composition approaches. Artist-researchers were recruited with music and AI skills and tasked with composing a piece of music using an AI tool of their choice. The chapter contributes six first-person accounts from their practice, gathered through a novel data collection approach using *reflection boards*, where participants were asked to pause and reflect back on screenshots of their composing after every hour. The first-person accounts offer rich descriptions of a plurality of AI-based music composition practices, which could inform others' AI music making. The following chapter analyses the RiCEv2 measures captured throughout this study to systematically show different characterisations of how reflection occurs across the artist-researchers' AI-based music composition.

Chapter 8

RiCEv2 Analysis of Artist-Researchers' AI-based Music Composition

The previous chapter showed that there is a plurality of ways that composers reflect across a range of AI music tools and composition approaches. To remind the reader, six artist-researchers were asked to pause and reflect after each hour of their composition process and write first-person accounts on their experience. This chapter analyses the Reflection in Creative Experience Version 2 (RiCEv2) measures collected after each hour of each composer's composition process. It shows that reflection is characterised by temporal patterns in the Artifical Intelligence (AI)-based music composition process.

The chapter is organised as follows. Section 8.1 describes the analysis approach for the RiCEv2 data. The findings are presented in Section 8.2, followed by a discussion of the findings in context with the previous chapter and related literature in Section 8.3.

8.1 Data Analysis Method

Descriptive statistics and visualisations are used to analyse the RiCEv2 measures across the participants' composition processes. More advanced modelling is not conducted given the small sample of seven artist-researchers.

The descriptive statistics provide a summary of the features of a sample (Müller et al., 2014). This includes measures of central tendency (mean and median). For context, these are compared to RiCEv2 scores calculated from the data collected for CSTs in Chapter 3. This provides a point of comparison for interpreting the RiCEv2 scores.

To characterise temporal patterns in reflection, visualisations are presented (Dix, 2020, pg. 124). They show the RiCEv2 scores changing over time by plotting the scores on the y-axis against time on the x-axis. Distinct shapes and colours are used to identify each participant. The benefit of visualising each individual is that characterisations of reflection can be identified within the heterogeneous data.

8.2 Findings

This section first presents the visualisations of the RiCEv2 measures across the composition process. This is followed by comparing this study's RiCEv2 scores and RiCEv2 scores calculated from the Chapter 3 data.

8.2.1 Visualisations

Figures 8.1 through 8.3 show plots for the three RiCEv2 metrics retrospectively reported by the participants after each hour of composing: reflection-on-process (Figure 8.1), reflection-on-self (Figure 8.2) and reflection-through-experimentation (Figure 8.3). To illustrate the changes in reflection over time, a plot of curved mean average trend lines for subsets of participants is visualised in Figure 8.1 and Figure 8.2. A linear trend line is plotted in Figure 8.3. The trends are based on patterns observed by the researcher across participants' RiCEv2 scores.

For **reflection-on-process** (Figure 8.1), P1, P3 and P4 show peaks in their scores in Session 1 and Session 3 (Trend A). This shows that they considered different directions to take their music at the start of their composing and before finalising their compositions. In contrast, P2, P5 and P6's scores generally decreased after Session 2 (Trend B). This shows that these participants considered alternative directions for their music early in their composition process. It is also noted that all the reflection-on-process scores are high, with none falling below four. Reflection on where to take a piece of music thus occurred to some extent throughout the music composition process. P1 scored the lowest across all sessions.

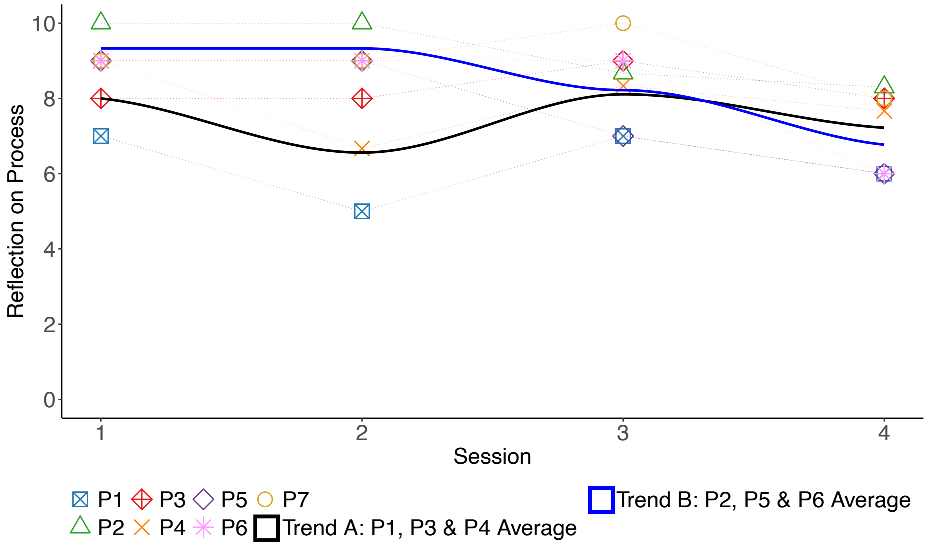


Figure 8.1: Artist-researchers' reflection-on-process scores for each session of their composing.

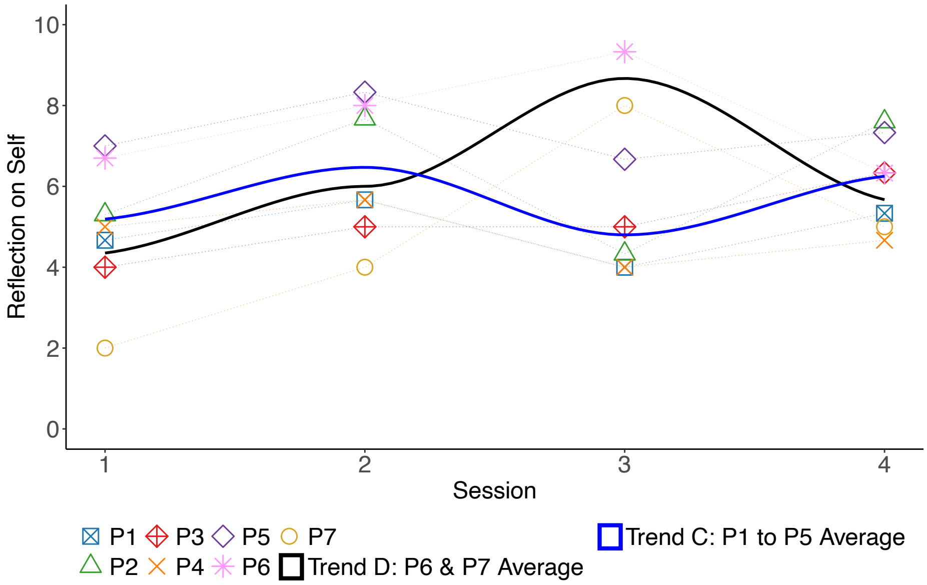


Figure 8.2: Artist-researchers' reflection-on-self scores from RiCEv2 for each session of their composing.

For **reflection-on-self** (Figure 8.2), participants 1 through 5 showed peaks in Session 2 and Session 4 (Trend C). This demonstrates a temporal fluctuation in how the participants reflect on their personal experiences. However, P6 and P7 gradually increase to a peak in Session 3 (Trend D). This shows that they reflected-on-self at different points in their composition process to the other participants.

For **reflection-through-experimentation** (Figure 8.3), a decline is observed over time (Trend E). This is partly driven by the outlier P3 in Session 4 (who mostly took on production duties at this moment), but there is also a clear decline across sessions from P2 and P5. In Session 2, it is noted that scores converge, and then diverge by Session 3. P6 and P7 annotated on the plot show high scores in Session 3, whilst P1, P3 and P4 annotated on the plot show low scores in Session 3. This shows that the changes in participants' activity between Session 2 and Session 3 increased or reduced reflection-through-experimentation.

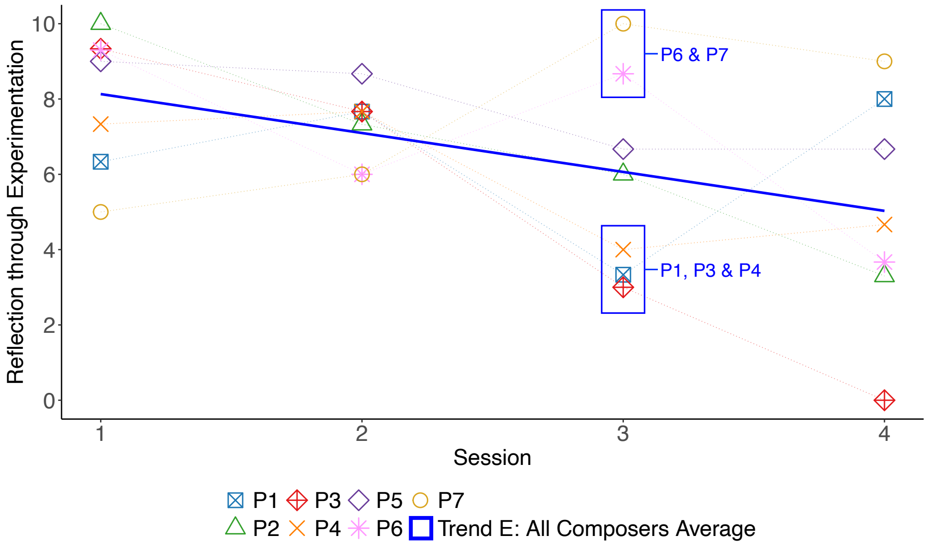


Figure 8.3: Artist-researchers' reflection-through-experimentation scores from RiCEv2 for each session of their composing.

8.2.2 Comparison of RiCEv2 Scores

This section compares the average RiCEv2 scores across all sessions with RiCEv2 scores calculated from the Chapter 3 data. The Chapter 3 was revisited, and RiCEv2 scores were calculated for CSTs, including: Photoshop, Word and some DAWs (such as Cubase, Ableton and Logic Pro).

Table 8.1 shows the reflection scores. This chapter's study participants have a RiCEv2 score equal to the Chapter 3 score for all CSTs (mean=6.8). The participant's reflection-on-process scores are higher than the Chapter 3 data. Reflection-on-process is lower for the DAW scores (mean=7.6) than this study's data (mean=8.1). Reflection-through-experimentation is higher for the DAW scores (mean=7.1) than this study's data (mean=6.6).

Table 8.1: RiCEv2 scores calculated from Chapter 3's data (top) and for this chapter's study (bottom). The DAWs include Cubase (n=2), Garageband (n=2), Ableton (n=2), Logic and FL Studio.

Dataset	Process	Self	Experiment	RiCEv2
Chapter 3 data				
All CSTs (n=300)	7.4	6.1	6.9	6.8
MS Word Subset (n=43)	7.2	6.4	6.6	6.7
Photoshop Subset (n=42)	7.4	5.9	7.1	6.8
Visual Studio Subset (n=15)	8.0	7.1	7.3	7.5
DAWs Subset (n=8)	7.6	5.7	7.1	6.8
This chapter				
Participant average (n=7)	8.1	5.8	6.6	6.8

8.3 Discussion

This chapter showed that reflection is characterised by temporal patterns in RiCEv2 scores across a variety of AI-based music composition processes (see Figures 8.1, 8.2, and 8.3). The main findings are shown in Table 8.2. This section triangulates these findings with the qualitative findings of the previous chapter, and discusses the findings in relation to relevant literature. The findings that contribute to new knowledge on AI music making are discussed in Section 8.3.1. Section 8.3.2 discusses findings similar to current research. The study limitations are outlined in Section 8.3.3.

Table 8.2: Summary of Chapter 8's main findings.

Main Finding	Location
Reflection is characterised in AI-based music composition as a push and pull between reflection-on-process (when curating AIGC in real time) and reflection-on-self (when arranging already curated AIGC).	Figure 8.4 (pg. 149)
Reflection is characterised as including a high amount of reflection on process in AI-based music composition. — Self-reflection is less prominent. — Reflection-through-experimentation decreases over time.	§8.2.1 §8.2.2

8.3.1 Novel Patterns

The observed patterns in Figures 8.1, 8.2 and 8.3 are explained through examining the first-person accounts in Chapter 7. For instance, it is shown that participants were listening to music in real-time when reflection-on-process is high (Figure 8.1). This interpretation is based on P1 who was improvising to select AIGC in Session 3, and P3 and P4, who were improvising to select musical layers in Session 3.

Reflection-on-process remained high across all sessions. It was also higher than the other CSTs in Table 8.1. This demonstrates that participants frequently reflected on alternative directions to take their music in the AI-based music composition process. Indeed, Figure 8.1 shows high reflection-on-process scores across all sessions.

Higher reflection-on-self scores (Figure 8.1) occurred when participants were *arranging* their AIGC. Notably, P6's high reflection-on-self score in Session 3 occurs alongside their description of self-reflection in Theme 1 (Reflection on Past Instincts); they reflected on the instinctual decisions that their *past self* had created in the previous session, learning about themselves by analysing their choices retrospectively. There is also evidence of Candy's (2019) reflection-at-a-distance where P6 was purposefully distancing themselves from their earlier decisions to assess their work more objectively.

A trade-off relationship is characterised by the observations of reflection-on-process and reflection-on-self, as visualised in Figure 8.4. When selecting AI

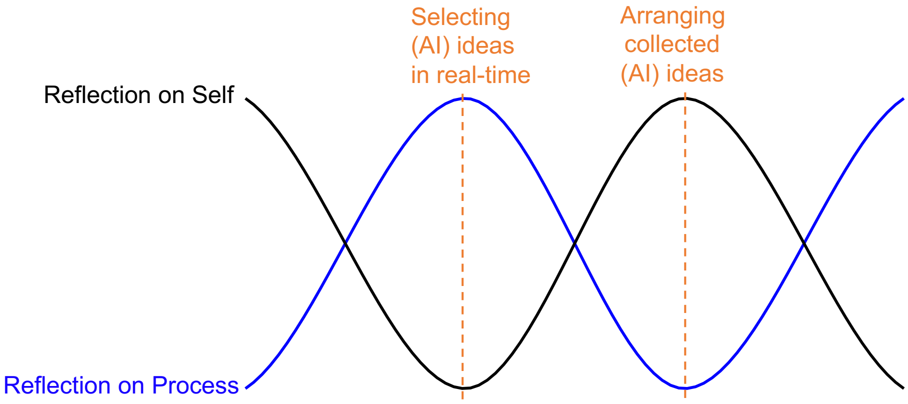


Figure 8.4: Model of the trade-off relationship between reflection-on-process and reflection-on-self. People selected ideas whilst listening in real-time and arranged their ideas after curation.

outputs and listening in real-time, participants reflected on future directions to take their music. They then reflected on what their music means to them when combining AIGC. This model supports definitions of reflection as moments where people sit back in contemplation (Moon, 2013; Wilson et al., 2023), and descriptions of its push-and-pull with moments of more instinctual reflection-in-the-moment (Candy, 2019).

There is evidence that setting up live-feedback to generate initial material by reflecting-in-the-moment (using instincts) was conducive to scaffolding opportunities for both reflection-on-process. In Theme 1 (Reflection on Past Instincts), participants described listening to AIGC in real-time as similar to their experience of improvising and playing in bands. Interacting with AI in this way is familiar to many artist-researchers and, therefore, appropriate for musical contexts. Whereas previous research showed that researchers adapted AI for their own practice (Gioti et al., 2022; Sturm et al., 2019), this study shows that this occurs alongside reflection-on-process.

8.3.2 Patterns Similar to Current Research

The findings from Theme 2 (Reflection on Direction and Surprises) support the high reflection-on-process scores. Participants would leverage surprising outputs to change the direction of their music. This confirms that Caramiaux and Fdili Alaoui's (2022) findings that AI-artists leveraged surprising outputs in their creative process applies to AI-based music composition. However, reflection-on-surprise cannot be attributed exclusively to AI tools. This is because P5 reflected on a surprising cowbell sample found in their live-coding environment, without using their AI. Reflection on surprising glitches has also been shown to occur in non-AI composition practices (Dannemann et al., 2023; Hamilton, 2020). Overall, the findings demonstrate that reflection-on-surprises (Candy, 2019), with or without AI, leads to speculation on the future and how people consider new directions to take their art.

Reflection-through-experimentation generally decreased over time, with participants converging in Session 2 and diverging in Session 3 (Figure 8.3). Given the findings in Theme 6 (Reflection on Technical Challenges), this shows that participants needed to first reflect on technical issues (cf. Candy's (2019) reflection-for-action) in Session 1 to later experiment with AIGC. Furthermore, some participants' reflection-through-experimentation decreased in Session 3 (in Figure 8.3) as they had already curated and decided how to organise their AIGC. Thus, they no longer needed to experiment.

8.3.3 Limitations

As emphasised in the previous chapter, there is variation in the data collected. For example, the composers have different practices and approaches to music composition, work in different genres, and use different AI tools. The set of AI tools used is heterogeneous and nested within a complex ecosystem of software and hardware (McGarry et al., 2017); a conflation between various types of tools in the findings is acknowledged. This research is positioned as generative and helpful in suggesting directions for future work.

The variation in data collection inhibits statistical analysis or generalisation from the findings. A clear opportunity exists to design more controlled A/B tests to untangle these factors in future work. For example, the comparison between DAW and AI measures in Table 8.1 demonstrates that AI encouraged more reflection-on-process. A more controlled A/B study design could confirm this, whilst accounting for confounding variables.

The first-person accounts helped explain patterns observed across the RiCEv2 measures, and brought more nuanced insights to the thematic analysis findings in Chapter 7. For example, references to researchers' inspirations and influences in the first-person accounts (see Theme 5) are not captured by the RiCEv2 questionnaire, nor other questionnaire measures typically used in creative HCI research (Cherry & Latulipe, 2014; Kerne et al., 2013). Investigating the impact of these more artistic influences and how to capture nuance in creative HCI and AI contexts is addressable in future work.

8.4 Conclusion

This chapter showed that reflection is characterised by temporal patterns across a variety of AI-based music composition processes. A characterisation of reflection in AI-based music composition is presented, which shows that reflection-on-process occurs when curating AIGC, whilst reflection-on-self occurs when organising the curated AIGC. This provides value for musical AI-based CST users, who could mimic these interaction patterns to spark different types of reflection in their own practice. The following part of this thesis evaluates a new AI-based CST for music composition using RiCEv2, where reflection support is a central design goal, to characterise reflection in a more homogeneous context.

Part III

Designing a New AI Music Composition Tool for Reflection

Chapter 9

Iterative Design of wAIve

The previous chapters showed that there are a variety of reflection patterns in existing Artificial Intelligence (AI)-based music composition tools. This motivates the development of a new AI-based Creativity Support Tool (CST) for music composition, with reflection support as a central design goal. This new tool focuses evaluation onto features designed to support reflection, rather than secondary concerns such as usability (Bryan-Kinns & Reed, 2023). It also limits heterogeneity, whilst still representing interactions common to AI-based music making.

This chapter describes the iterative design of the new tool, named wAIve. For four iterations each, participant pairs were recruited to offer critique and suggestions to emphasise reflection in wAIve's user experience. Each pair had different skills of data science, AI and music interfaces, and design. The chapter shows how they informed design features for wAIve which emphasise reflection.

The chapter is organised as follows. First, the iterative design method is described in Section 9.1. The findings of each iteration are presented chronologically, grouped by participant pairs' skills in Section 9.2. The chapter closes with a discussion on the iterative design process and wAIve's features in Section 9.3.

9.1 Method

Iterative design was used to identify features emphasising reflection in wAIve. Participants interacted with versions of wAIve over multiple sessions. Their evaluation of the interface was captured using semi-structured interviews. An iterative design approach was selected instead of a more linear approach because of the lack of a clear definition for reflection. With iteration, a mutual understanding of reflection in the context of wAIve could evolve between the participants and the researcher (Hartson & Hix, 1989).

The study was approved by Queen Mary University of London's ethics committee. Each participant was reimbursed with a £25 Amazon voucher.

9.1.1 Design Goals

Three initial design goals were set to develop wAIve:

- **Goal 1:** To encourage opportunities to reflect whilst making music.
- **Goal 2:** To not include features too obtrusive as to break engagement.
- **Goal 3:** To provide an intuitive interface where non-musicians create music of an acceptable quality.

Goals 1 and 2 were purposefully constructed to be in contention with each other. The contention encourages a push-and-pull between reflection and engagement, and critical thinking on their interplay. Goal 3 ensures that wAIve produces acceptable quality music to maintain motivation and engagement.

9.1.2 Participants

A call for participation was sent to Queen Mary University of London's e-mail lists for PhD and Master's students in the Electronic Engineering and Computer Science department. Academics were also contacted to identify Master's students or final-year undergraduate students who are interested in participating and can provide reflective, critical feedback. University students were targeted to ensure they had formal training in their areas of expertise.

Three pairs of participants were recruited in total. Each pair had different skills because human-AI interaction design requires understandings from various disciplinary perspectives (Yang et al., 2020). The pair's skills were in: data science, AI and music interfaces, and design. These different skills were selected to gather data on non-musicians' and musicians' perspectives, AI features, music and interaction aspects, and wAIve's look and feel. Participants were paired by their discipline to focus discussion on the development of wAIve and not on how people from different disciplines think about AI. Recruitment was stopped after three pairs because each pair's perspectives covered most aspects of wAIve's interaction design. Herein, the participants are referred to as the data scientists (DS1 and DS2), AI musicians (AIM1 and AIM2) and designers (DE1 and DE2).

The participants' expertise and demographics were captured using a pre-test questionnaire. Goldsmiths MSI (Müllensiefen et al., 2014) scores assessed musical sophistication; for context, the UK national average MSI is 81.6. SRIS scores (Grant et al., 2002) were gathered to assess reflective capacity; for context, the average SRIS score for users of CSTs from Chapter 3 is 4.5.

Table 9.1: Goldsmiths MSI and SRIS scores for wAIve's design study. Participants are labelled by their expertise where DS = data scientists, AIM = AI and music interfaces, and DE = design.

Scale	DS1	DS2	AIM1	AIM2	DE1	DE1
MSI Score	**62**	**51**	**110**	**95**	**51**	**39**
Mean Active Engagement	2.8	2.8	6.3	6.0	2.8	3.5
Mean Perceptual Ability	5.0	3.5	6.5	6.5	5.0	3.5
Mean Musical Training	2.2	1.4	5.8	4.4	1.0	1.0
Mean Singing Ability	4.2	3.8	6.2	5.0	3.5	1.3
SRIS Mean	**4.9**	**4.8**	**5.7**	**4.8**	**5.2**	**4.3**
Mean Engagement	3.8	4.2	4.5	4.2	4.0	4.2
Mean Need	3.5	4.2	4.3	3.2	3.5	2.5
Mean Insights	3.8	2.8	4.3	3.6	4.3	3.1

Table 9.1 shows the MSI and SRIS scores. AIM1 and AIM2 have MSI scores above the UK average, whilst the other pairs are below average. The SRIS scores across participants are close to but higher than those of the CST users in Chapter 3. This is interpreted as the sample showing reflective skills representative of other CST users and thus does not bias the study's findings.

9.1.3 Procedure

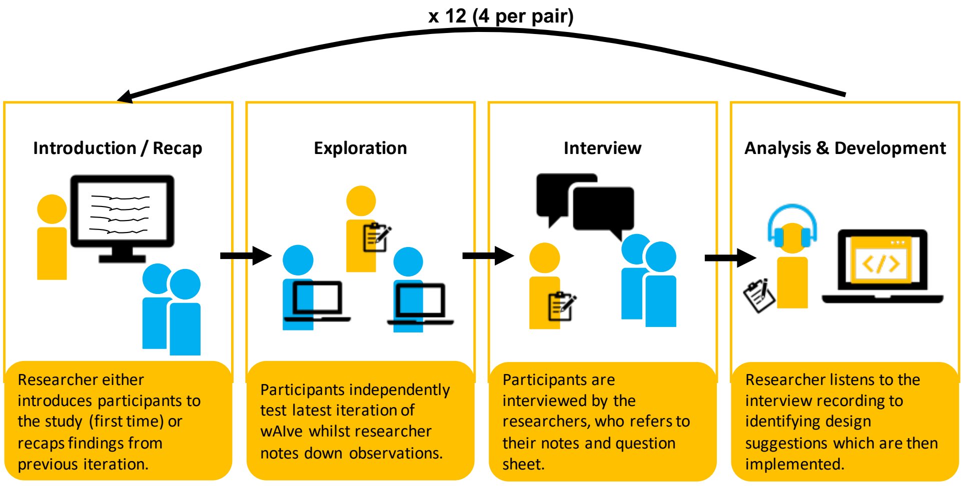

x 12 (4 per pair)
Introduction / Recap
Exploration
Interview
Analysis & Development
Researcher either introduces participants to the study (first time) or recaps findings from previous iteration.
Participants independently test latest iteration of wAIve whilst researcher notes down observations.
Participants are interviewed by the researchers, who refers to their notes and question sheet.
Researcher listens to the interview recording to identifying design suggestions which are then implemented.

Figure 9.1: Overview of the iterative design process procedure for wAIve's design study. The researcher is orange. The participants are blue.

Figure 9.1 shows an overview of the design process used for wAIve. After the **initial design** phase (see Section 9.2.1), pairs of participants met with the researcher (who is also the thesis author) for four sessions. Each session was either in-person or on Zoom[1] to lessen the burden to attend sessions. Twelve sessions were completed (four sessions multiplied by three pairs of participants). The sessions lasted 30 minutes each and were audio recorded. The sessions ran as follows:

1. **Introduction or Recap:** In the first session with a pair of participants, pairs completed the pre-test questionnaire (see Section 9.1.2). They were then given a demo of wAIve. Next, ground rules (see the appendix) were set to encourage critique (Braun & Clarke, 2013), and pairs were introduced to the design goals in Section 9.1.1. For the other three sessions with a pair of participants, the researcher reiterated the design goals and summarised findings from the previous iteration. This refreshed participants' memory and allowed time to correct misinterpretations.

2. **Task:** Participants composed music using wAIve for ten minutes. The timing was taken from Ford and Nash (2020), which, although

[1] https://zoom.us/

it presents a different context, served as a useful starting point for this study. The instruction was: "to freely compose a piece of music with wAIve". The task was purposefully open-ended to explore reflection in a creative experience with typical characteristics (Kerne et al., 2013). The researcher noted observations throughout.

3. **Interview:** The researcher interviewed each participant. A semi-structured approach was used to probe spontaneous interactions and to discuss ideas at the forefront of participants' minds. The questions asked in interviews on engagement by Wu (2018) were taken as a starting point. Further questions were added to match the thesis's context as iterations progressed. The questions are shown in Table 9.2. Discussion was not limited to these questions; the researcher also asked questions based on their observations.

9.1.4 Data Analysis Method

After each session, the researcher reviewed the recordings and identified design ideas, as follows. The interviews were transcribed as described in Section 7.1.5. Next, the researcher highlighted possible design suggestions, feature requests, and critiques mentioned by the participants. This lightweight approach was used instead of thematic analysis (see Section 7.1.5) because of the limited time between design iterations. Design ideas were then implemented into wAIve for the next iteration.

9.2 Design Iterations

This section documents the main discussion points from each design iteration, using supporting quotes from participants. The subsections below describe the following: the initial design of wAIve (Section 9.2.1), the iterations with each pair of participants (Section 9.2.2 to 9.2.4), and the final prototype of wAIve (Section 9.2.5).

9.2.1 Initial Design

An initial design was developed for wAIve starting with the goals in Section 9.1.1. The user interface is shown in Figure 9.2. Existing CSTs inspired the design. This ensures that wAIve would be representative of other mu-

Table 9.2: Interview questions used as the basis for the semi-structured interviews in wAIve's design study. Questions marked with † were added as iterations progressed.

Opening Questions
- What are your initial thoughts on the interface?
- Did any features stand out to you as supporting reflection/engagement?†
- At what moments did you experience reflection?†
Design Goal 1: Reflection
- Did you find the interface encouraged you to reflect on your music making?
- How did you reflect on your music whilst using the interface?
- Were there any moments that prompted you to stop to think about your music making?
- Did any parts of the interface encourage you to reflect?
- What different kinds of reflection did you experience?†
Design Goal 2: Engagement
- Were there any points where you felt annoyed?
- Were there any points where you were frustrated?
- Were you distracted by the interface at any points?
- Was anything jolting or off-putting?†
Design Goal 3: Usability
- Do you think that the interface was intuitive?
- Could you make quality-sounding music with the interface?
- Did the interface help you to create your music?
- What features of the interface helped to make your music?
- What features of the prototype would you improve so that the interface was easier to understand?
- Do you like the music that you created?
Probes
- Would you explain further?
- Would you give an example?
- How would you go about explaining this?
Closing Questions
- Of all the things we've talked about, what is most important to you?
- Any features that you think are missing?†
- Any general suggestions?†

sic CSTs to study common interactions. The CSTs are as follows. First, *Codetta* (Ford & Bryan-Kinns, 2022a; Ford & Nash, 2020; Ford et al., 2021), where draggable puzzle-shaped blocks of notes helped children write music quickly, cf. Goal 3. Second, *combinFormation* (Kerne et al., 2014), which finds images related to a user's image collection and organises these into related clusters around the edges of their collection. Based on their experience as CST designers, the researcher postulated that this approach was more conducive to the open-ended nature of music composition than alternatives such as sliders (Louie et al., 2020).

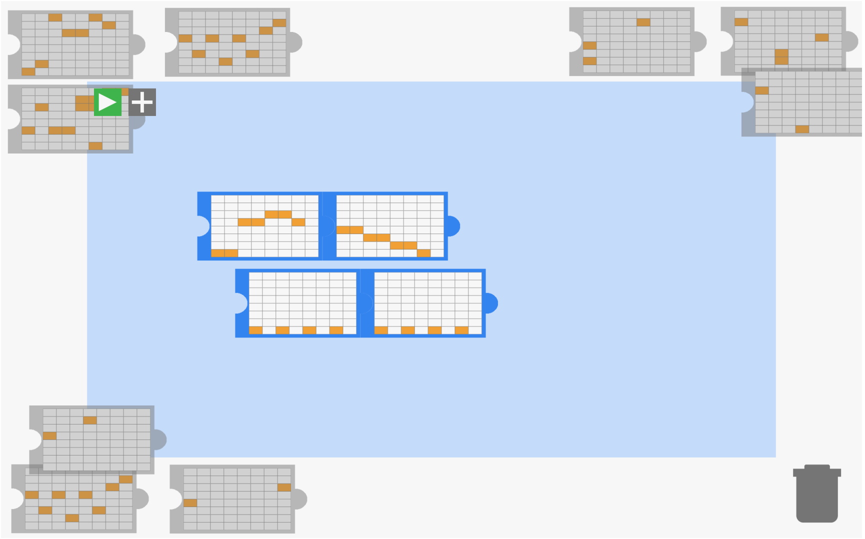


Figure 9.2: WAIve's initial design.

In wAIve, blocks of generated music were organised around a central workspace. The blocks are grouped based on the musical metrics of: pitch count, pitch range and average pitch interval. This metric approach was inspired by Banar and Colton's (2021) interface for composing with ChatGPT-styled outputs (see pg. 111). The selected musical metrics were also used in other music AI research (Banar and Colton, 2022; Yang and Lerch, 2020). The technical details are described in Section 10.2.2. A grid-based interface for each block was chosen instead of a music notation system under the assumption, based on the researchers' experience, that it would be more intuitive to non-musicians (Goal 3), yet still common to music software (Rossmy, 2022). This initial design was taken forward for testing, starting with the pair of data scientists.

9.2.2 Iterations 1 to 4 (Data Scientists)

The data scientists evaluated the first four iterations of wAIve. In iteration one, the data scientists found it difficult to reflect on similarities between their own music and AI generated blocks. DS2 said they were "unable to identify... similarity in one [... block] being with the other one". DS1 requested to highlight the blocks' notes to clarify similarities: "going up[...] means, you are going happier, so maybe a colour representing happy[...] and then if it's going down, it is sad".

For iteration two, colour patterns of notes were added to test if it became easier to reflect on similarities and differences between blocks. As shown in Figure 9.3, colours were tested across iterations for: ascending melodies (in green) and descending melodies (in purple). The author chose the colours because they wanted to reflect the data scientists' notion of "happy" and "sad" music, but avoid connotations of good and bad from colours such as green and red. Overall, the data scientists found the different colours helpful in reflecting on differences between blocks – for example, DS2 indicated that "with purple and green there was something different".

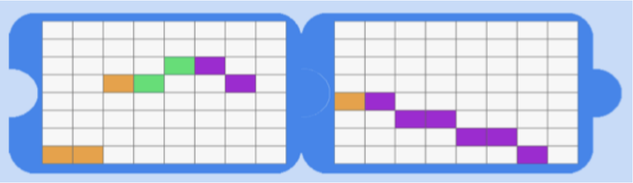


Figure 9.3: Note colours added to represent ascending and descending melodies in iteration 2 of wAIve's design.

However, the data scientists described placing similar coloured blocks together without considering how the music sounded. They did not reflect on the blocks' musical qualities (Goal 1). For example, DS2 said "I would match the colours in it". DS1 said "when it comes to decision making, I[...] rely on visualisation. I actually rely on the colours to decide whether I want it or not. It's not actually the sound.". In considering this, DS1 described how "the AI shouldn't lead us [to reflect], it should just help us".

A fading animation was added for iteration four to balance visualising note patterns using colours with encouraging reflection on musical quality. As shown in Figure 9.4, note colours would fade to orange after 25 seconds.

This was inspired by related work (An et al., 2019) where fading coloured beacons were used in a classroom to indicate whether students had recently had support. The intuition for wAIve was that gradually fading note colours would bring the benefit of identifying similar blocks using colours, and encourage reflection shortly after. However, DS1 said the fading was "a barrier for me, my listening is not waking up".

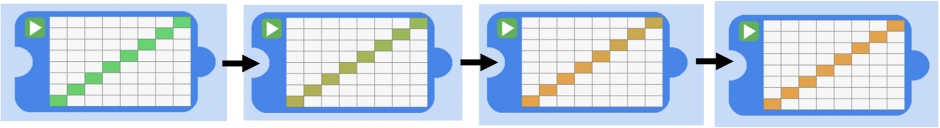


Figure 9.4: Storyboard of the fading note colours added in iteration 4 of wAIve's design.

Listening to blocks in different ways was considered important for usability (Goal 3). For example, DS1 said "I wish that I could just click on some button" to "listen to a segment and decide what I want to do about [it]". Across iterations one to four, various ways to listen to the blocks were thus implemented. This included adding small play buttons on each block and a second workspace, shown in Figure 9.5. These features meant that users could listen at three levels: i) an individual block, ii) all blocks in a workspace, or iii) all the blocks together. DS2 said that the different listening options "gives me an advantage to play the entire [... music...] differently". DS1 noted that these options helped them in "thinking about the entire process". This shows that the playback options supported reflection-on-process.

9.2.3 Iterations 5 to 8 (AI Musicians)

Before the iterations with the AI musicians, the note colours' fade out time was extended. The intuition was that participants would have more time on screen to consider patterns of colours. However, the AI musicians found this "confusing" (AIM2) instead of encouraging reflection. The feature was thus removed.

The AI musicians found the note colours difficult to interpret. AIM1 was "not completely sure why or what [the colours] represent". To support this, AIM2 requested that "blocks that are connected all take on one particular colour" (AIM2). Indeed, AIM2 suggested using different colours for the

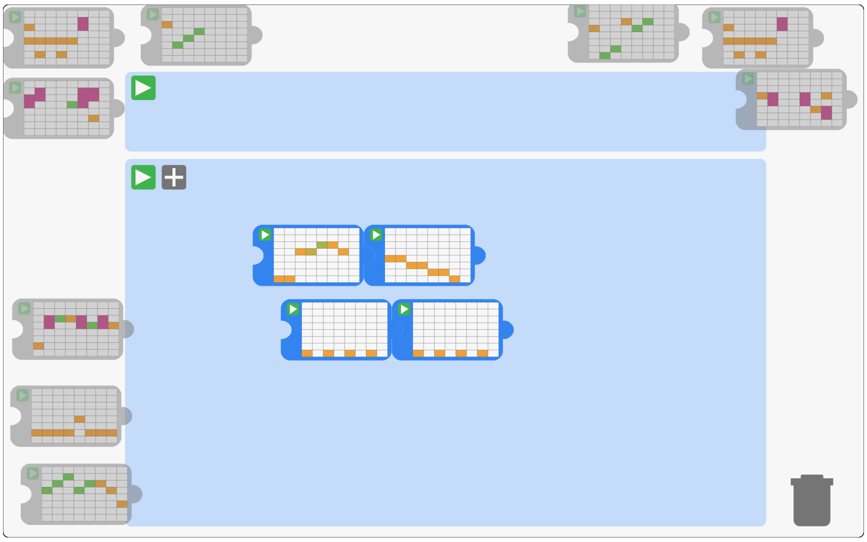


Figure 9.5: Iteration 4 of wAIve with an additional workspace and play buttons added to each block.

different workspaces, citing the music software Ableton[2], where "each track is a different colour, so you only need one look to know if you are looking at drums or synths or whichever".

AIM2 also requested to change to three workspaces, closer to a DAW layout (see Figure 9.6). This supported the design of wAIve as a tool representing interactions common to other AI music tools. AI generated blocks to the right of each timeline would imply "that they are supposed to be appended at the end" (AIM2). These initial evaluations of wAIve described above enhanced the AI musician's familiarity with wAIve and perception of its usability (Goal 3). They required a solid foundation before they could investigate more complex ideas on reflection in later iterations.

In iteration six, AIM1 identified that simpler AI algorithms to the music metric-based approach could better support reflection on "variations on existing blocks [or ...] putting things in opposite directions". The music metric-based algorithm was thus replaced with Google Magenta's MusicVAE (Roberts, Engel, et al., 2018) model in iteration seven. MusicVAE was able to generate a more direct match to the users' music in each timeline by identifying a latent representation for the users' music and then modifying it slightly to create music with minor differences (see technical details in Sec-

[2] https://www.ableton.com/

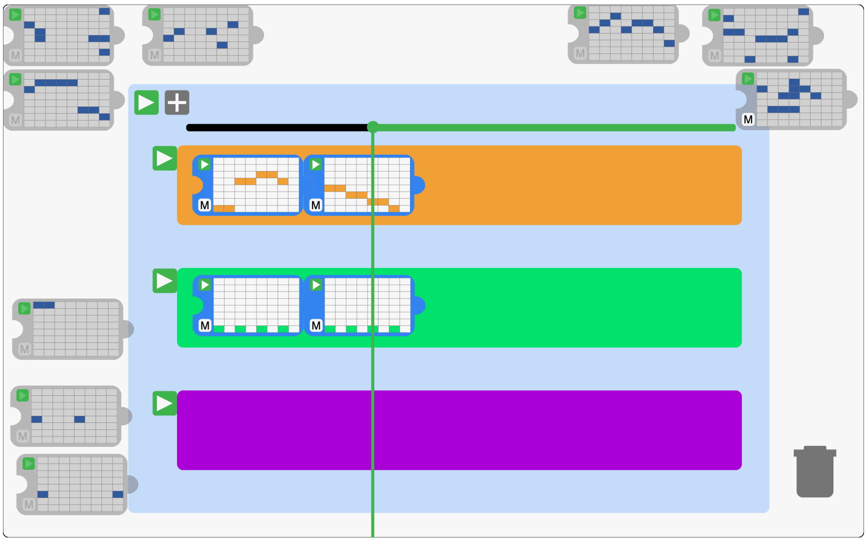


Figure 9.6: Iteration 6 of wAIve with multiple timelines. There is a different colour per timeline.

tion 10.2.1). As an established model, MusicVAE could also generate high quality music cf. Goal 3.

Initially, the MusicVAE outputs were seen as "mirrors of what was already there" (AIM2), yet "allowed you to manifest an idea much quicker, whereas before it was just random" (AIM1). The similarity was later modified to a "sweet spot" (AIM2), which reduced the degree of similarity between the generated sequence from MusicVAE and the input sequence.

The AI musicians, like the data scientists, found that the different playback options encouraged reflection on different perspectives. This is because "if you think about reflection on different levels – on a hierarchy, then – you can describe each of them individually" (AIM2). A system to flash the play buttons was developed to encourage a variety of playback. The play buttons that were clicked the least within 25 seconds started to blink (see Section 10.3.2). Various designs were tested in iterations seven and eight. First, bright red flashing was tested. However, AIM2 described these as "invasive" and AIM1 as "intense". The buttons were then changed to a lighter blue, closer to the background. Examples are shown in Figure 9.7.

Discussions on the flashing button feature raised the question of what kinds of reflection would be best in the context of music composition. AIM2 said

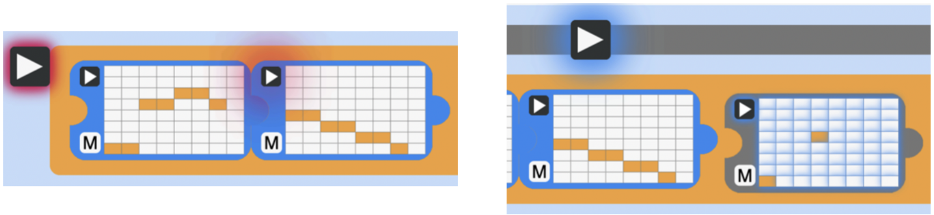


Figure 9.7: Flashing play buttons. Implemented in iterations 7 (left) and 8 (right) of wAIve's design.

that reflection when composing is not "necessarily cyclical and repetitive[...] if I'm happy with something that I've put down, how and when will I reflect on it?". AIM1 said "the way I reflect is by going okay, here is this composition, I have two parts that are playing at the same time... how do I make those two parts sound good together?" Whilst reflecting on different perspectives was encouraged by flashing different play buttons, these perspectives emphasise the importance of reflecting on new material and how it fits with the user's musical context, cf. reflection-on-process (Chapter 3).

Animations for the blocks to fly into the workspace were developed to test if they could help with reflection. AIM1 said "if [wAIve is] for someone who does not[...] know what is a good decision[...] making the blocks go in automatically[... they might] reflect and go, 'that sounds good?', 'that sounds bad?"'. The AI generated blocks implemented move to the end of a timeline (randomly chosen) and, if the block was not clicked, move to the bin after 25 seconds (see Section 10.3.1 for technical detail). The AI musicians found the flying blocks obtrusive but not annoying per se. For example, AIM2 liked that the moving blocks "come into your line of sight" and "allow you to focus on the music as a whole rather than [...] individual pieces".

9.2.4 Iterations 9 to 12 (Designers)

When the designers first tested wAIve, their initial thoughts turned to the colours of the blocks. For example, DE2 asked "What do the general colours mean?". The links between AI generated block colours and timeline colours were not obvious and caused confusion. To better demonstrate this link, animated curves were added, shown in Figure 9.8. Their design was inspired by a musical CST named Manhattan (Nash, 2014), which uses curves to show relationships between hidden user interface elements. The curves appeared on screen for five seconds whenever new AI generated blocks were added to wAIve's user interface (see Section 10.3.3 for technical details). These were seen as "nice" (DE1) and "helpful" (DE2). DE1 said the curves "occasionally pop up, so it wasn't annoying" (DE1). These also "prompted [DE1] to look to the side [and] add some pieces of music" (DE1), which they would not have reflected on otherwise.

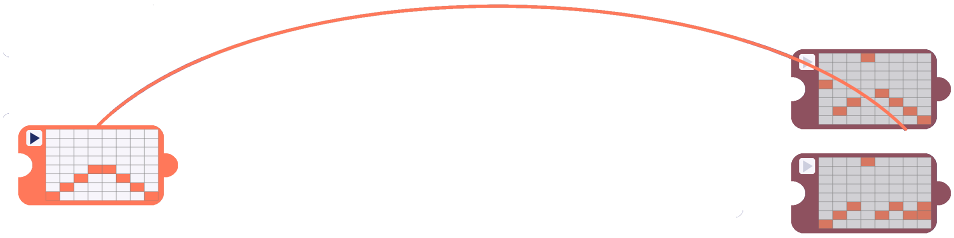


Figure 9.8: Animated curves. These appear between AI-generated blocks (right) and blocks in a timeline (left) when new generations are added.

The flying blocks caused confusion upon initial testing but later led to a moment of reflection. After DE1 said "I was so confused; they were just all moving around like someone else who was using it", DE2 said "I had the same thing." Yet, "when the things were moving around[...] it was prompting me to think, why are they giving me these suggestions?". This shows that the flying AI animations prompted reflection on the AI and how wAIve worked (Goal 1). Thus, they were brought forward to the final design for wAIve.

Various instrument sounds were implemented in iteration eleven because the designers assumed this was what the colour coding of timelines indicated: "I thought it was different instruments or differently generated sounds" (DE2). Previously, all notes were played with a square wave synthesiser sound. In iteration eleven, the piano, clarinet and cello were chosen for their distinct timbres. The designers liked these; however, they requested "the iconic instruments" (DE2). Both enthusiastically requested "drums", speaking at the same time. The instruments were changed to guitar, bass and drums for iteration twelve.

New colour pallets for wAIve were tested in iterations eleven (see Figure 9.9) and twelve (see Figure 9.10). The designers found that darker colours made the user interface "more clear" (DE2) and closer to professional music software. The original colour scheme was "more child-like" (DE2). However, their assessment of the design in iteration eleven was that it was too dark, leading to the final colour scheme in iteration twelve. The flashing for the play buttons was also modified to complement the new colour schemes, moving from the intense colours of previous iterations to more neutral colours.

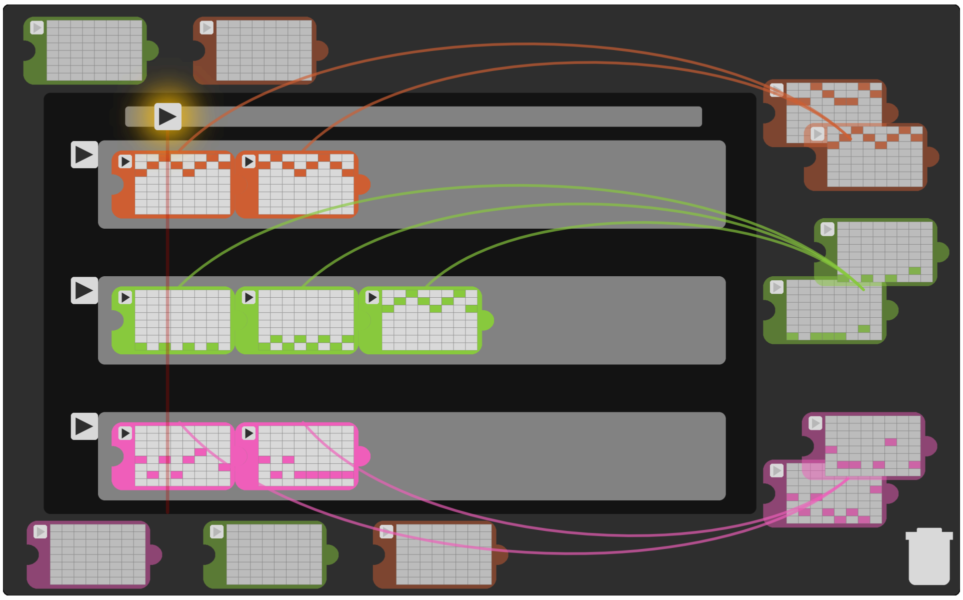


Figure 9.9: Iteration 10 of wAIve, with the dark colour scheme guided by the designers.

9.2.5 Final Prototype

The previous sections described how participant pairs with different disciplinary perspectives informed features for wAIve. Here, the final prototype of wAIve and its interaction is described. Technical implementation details are in Chapter 10.

The final iteration of wAIve is shown in Figure 9.10. It operates as follows. Blocks are connected by dragging their left side to the right side of another block. Left-side dragging of blocks moves stacks of connected blocks; right-side dragging disconnects blocks. The grid buttons within each block add

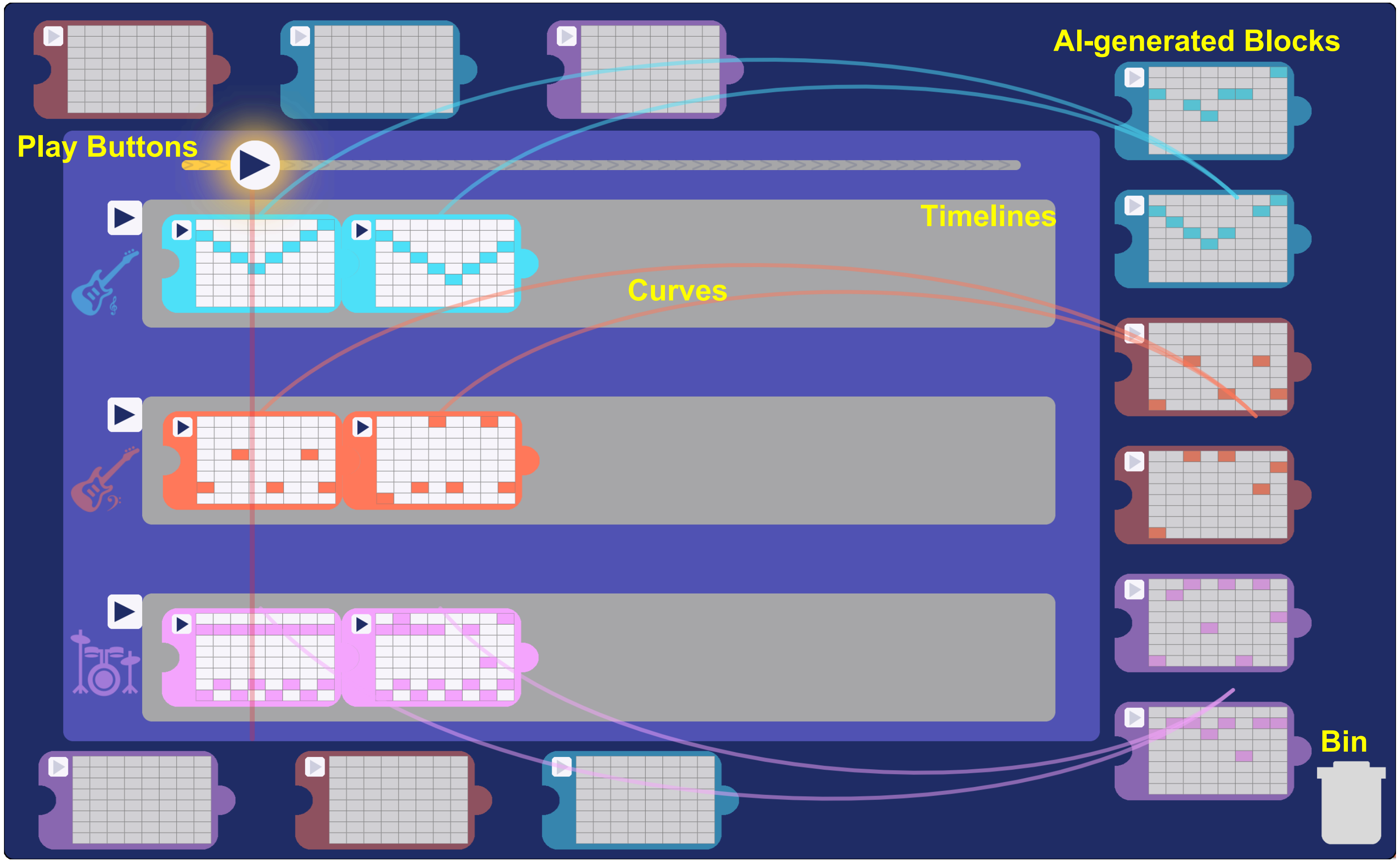


Figure 9.10: WAIve's final design. The features are labelled in yellow text. The text is not part of the user interface. WAIve can be tested online at https://codetta.codes/wAIve/.

notes. The play buttons on blocks play the notes in that block. The play buttons to the left of the grey rectangles, called timelines, play all the blocks in the area sequentially. The top play button plays all blocks together across timelines. Blocks can also be placed in the bin to delete them.

Every 25 seconds, wAIve oscillates between flashing play buttons and loading AI generated music, and moving an AI generation randomly to the right of a block stack. AI generated music is created by inputting the music in a timeline to Google's MusicVAE model (Roberts, Engel, et al., 2018). MusicVAE outputs similar, but not the same, music. The flying block motion stops on click; otherwise, after 25 seconds, it will move to the bin. Likewise, the play buttons stop flashing when clicking or after 25 seconds. Curves appear on the screen briefly between the timelines and AI generated blocks when blocks are first added.

9.3 Discussion

Table 9.3: Summary of Chapter 9's main findings.

Main Finding	Location
Reflection is characterised as occurring when people listen to their music from different perspectives. For example, listening to the piece as a whole encourages reflection-on-process.	§9.2.2 §9.2.3
Reflection is characterised as supported when people are familiar with an interface. For example, AI musicians requested that wAIve's design be closer to a DAW.	§9.2.3
Reflection is characterised as not occurring when participants can rely on visuals in lieu of listening.	§9.2.2
Reflection in AI-based music composition is characterised as occurring when people observe an AI to understand its actions.	§9.2.4

This chapter describes the iterative design of wAIve, with emphasising reflection as its central design goal (see pg. 153). The main findings are shown in Table 9.3. Through an iterative design process, several features were implemented. These features included: AI generated music that was similar but not too similar to the user's music, playback of music from different perspectives, flashing of play buttons that are sparingly used, and flying AI generated block animations. Below discusses the iterative design process and wAIve's features in context with related literature.

9.3.1 Iterative Design Process

A similar pattern was observed across the four iterations of each participant pair. In the first sessions, participants gave initial impressions focused on usability concerns (Goal 3). For example, the AI musicians asked to change wAIve's layout to a more familiar design. The designers also asked questions on how the AI and play buttons worked. It was not until the second sessions that participants offered divergent ideas related to reflection. For instance, less typical features such as fading of note colours, flashing play buttons or animating blocks to fly around the screen did not surface until the later iterations. In the final sessions, participants tested their divergent ideas.

This pattern of idea development aligns with existing models of the design process (see Section 2.1). For example, the generation of new ideas followed by their validation maps to the *develop* and *deliver* phases of the Double Diamond model[3]. The initial discussions on usability, laying the groundwork for more divergent ideas on reflection, also map to Sanders and Stappers's (2008) "fuzzy front end". The design space became clearer once participants were familiar with wAIve's interaction and understood the design problem. Participants also moved through stages of engagement as they became more familiar with wAIve, cf. Bilda et al. (2008). Future work could examine how to reach divergent ideas more quickly to test a wider range of ideas on reflection in less time.

The participants gave feedback on a programmed CST as opposed to a mock-up interface. The advantage of this was that participants commented directly on their interaction with real-world generative AI tools. Human-AI interaction design research (Yang et al., 2020) has more frequently used pretend prototypes. For example, in the Wizard of Oz procedure, users interact with a pretend interface, and a researcher acts as the AI (see Bellingham (2022) and Thelle and Fiebrink (2022)). By evaluating direct interaction with wAIve, there is a higher probability that the participants' insights will translate to similar contexts.

However, it was difficult to implement complex ideas into wAIve's code in the short time between iterations. Use of mock-up interfaces before pro-

[3] https://www.designcouncil.org.uk/our-resources/the-double-diamond/

gramming would make it easier to develop complex interactions. A cyclical development process also meant that updating the system required an understanding of programming, limiting the opportunity for more direct collaboration on wAIve's design with participants. For instance, the designers' limited knowledge of code meant they could not directly experiment with wAIve's colour palette. If they could have, this would have avoided the back-and-forth between different colour scheme designs in iterations eleven and twelve.

Each participant pair gave different perspectives on wAIve and its interaction. The AI musicians gave insights that led to the most prominent changes. For example, the AI musicians requested three timelines, the flying AI animations and flashing play buttons. Some insights reflected their expertise in music, such as AIM2's request to organise wAIve to be more similar to a DAW. These changes supported the design of wAIve as a tool for studying interactions common to AI-based musical CSTs such as Hunt et al. (2020) or Tchemeube et al. (2022). However, more creative or divergent ideas were harder to envision. In Boden's (1991) terms, the AI musicians showed exploratory creativity and not transformational creativity. Similarly, the data scientists' request for colour coding notes relates to their expertise, similar to identifying patterns in data with visualisations (Zaki & Meira Jr, 2020). These findings show that familiarity with wAIve's user interface, which is closer to the participants' domain, helped to encourage reflection.

9.3.2 WAIve and its Features

The development of different features for wAIve led to discussions on how reflection is characterised when composing music with AIGC. For example, colour coding notes led the data scientists to realise that they mainly reflected on colours, not how the notes sounded. For example, DS1 said "when it comes to decision making, I[...] actually rely on the colours". They also described how this was not the effect they thought the AI should have; instead, the AI "shouldn't lead us, it should just help us. This connects with the debates around how an AI should be given agency over the creative process (Amershi et al., 2019; Boden & Edmonds, 2009; Lewis, 2023; Louie et al., 2020; Wilson et al., 2023; Xambó, 2022). DS1's position is to leave autonomy to the user.

The AI musicians also discussed their reflection when composing. AIM2 said their process was not “necessarily cyclical” and did not consider previous material. This contrasts models of the reflection process where people reflect back (Boud et al., 1985; Kolb, 1984). AIM1 found they reflected on “parts that are playing at the same time”. This confirms that characterisations of reflection which include comparisons (Bentvelzen et al., 2022; Boud et al., 1985; Kolb, 1984) and decisions in-the-making-moment (Candy, 2019; Schön, 1983) occurred in this study’s AI-based music composition process.

The AI musicians found value in reflecting by listening to the composition in different ways. AIM2 described this as “a hierarchy” of listening at different levels. To encourage more diversity in listening, the flashing of play button options brought attention to options that participants otherwise might not have considered. This contrasts reflective CST designs, which offer different visualisations to show different perspectives (Belakova & Mackay, 2021; Hoque et al., 2022; Sterman et al., 2023). Different ways to *listen* were more important for the music composition context. This is supported by comments from the participants that colour coding was confusing, for example, DE2 asking “what do the general colours mean?” and AIM1 stating they were “not completely sure why or what [the colours] represent”. This is also supported by research emphasising the role of listening in music making (Nash & Blackwell, 2012).

In the AI musicians’ third iteration, wAIve’s AI was switched from the music metric-based algorithm (Banar & Colton, 2021) to inputting the user’s own music to Google Magenta’s MusicVAE (Roberts, Engel, et al., 2018). MusicVAE produced similar music by manipulating small changes in its latent space. Similarity between the AI and users’ music was a discussion point despite not being an explicit design goal, indicating its relevance to reflection or engagement (Goals 1 and 2). This corroborates findings from the continuator system (Addessi et al., 2015; Pachet, 2003) introduced in Section 6.2, where similar musical phrases returned to users in an improvisation context supported their engagement.

9.4 Conclusion

This chapter described the iterative design of wAIve: a new AI-based CST for music composition, with reflection support as a central design goal. Its features include AI generated flying blocks and multiple options for playback. Thus, wAIve enables the study of a homogenous sample of participants, with limited interactions to reduce variation in the study of users' composition processes and to focus on reflection, not other study goals. The following chapter gives details on the technical implementation of these features.

Chapter 10

Technical Implementation of wAIve

This chapter describes the technical implementation details of wAIve. This is to the level of detail such that other researchers can recreate the AI music generation and animations. It starts with a high-level overview of the main drawing loop, where the AI music generation and animations are executed (Section 10.1). The AI music generation algorithms and animations are then described in Section 10.2 and Section 10.3 respectively.

10.1 Overview

wAIve is a JavaScript application that uses two external libraries. The user interface is built using the P5.js library (McCarthy et al., 2015), designed to create images, animations, and interactive graphics. The Magenta.js library (Roberts, Hawthorne, & Simon, 2018) is used to perform inference with pre-trained music AI models and synthesise audio output.

P5.js renders graphics to a canvas at 60 frames per second. wAIve renders its animations and interface visuals within this loop. Figure 10.1 visualises the algorithm for triggering AI generation and rendering animations within the drawing loop. It first checks whether 25 seconds have elapsed, based on timings used to predict engagement in Ford and Bryan-Kinns (2022a)[1]. If elapsed, the code switches between selecting an AI block at random and triggering it to fly across the workspace, and generating new AI blocks with a curve animation and flashing underused play buttons.

[1] Although the findings are not directly generalisable due to wAIve's different context, the 25 seconds was used as a starting point. The iterative design process in the previous chapter demonstrated its appropriateness through user testing.

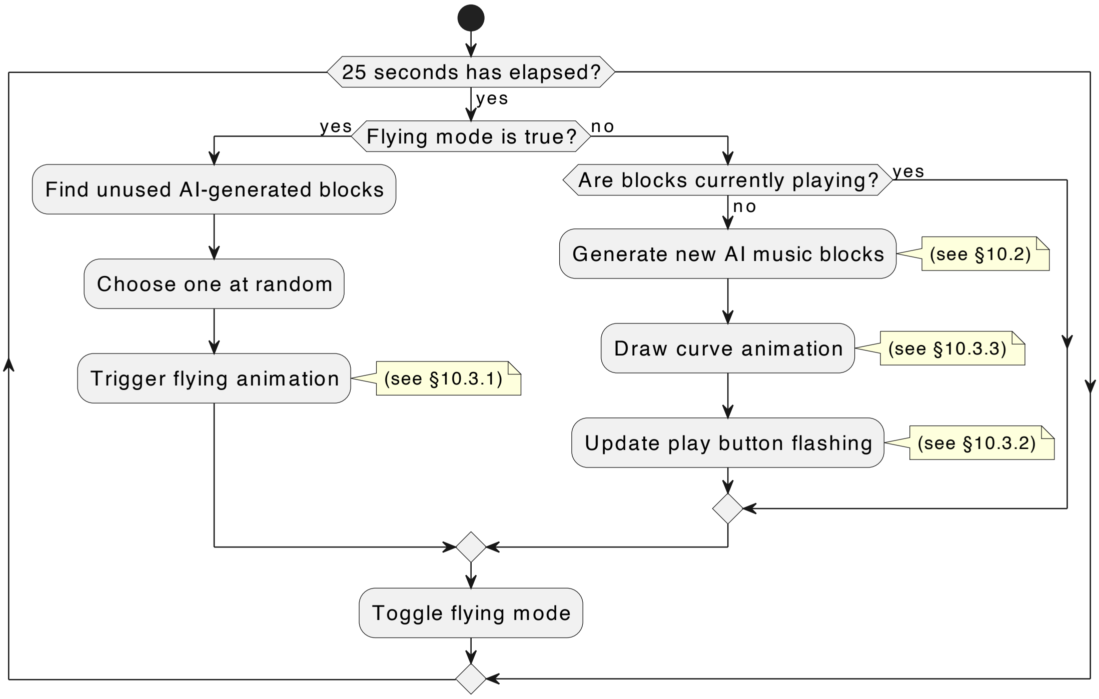


Figure 10.1: Flowchart visualising the drawing loop which triggers AI generations and animations in wAIve.

10.2 AI Music Generation

This section details the AI generation from MusicVAE (Roberts, Engel, et al., 2018) used in the final design of wAIve in Section 10.2.1. It then details the music metric-based algorithm that was used prior to iteration six (introduced on pg. 159 in Section 10.2.2).

10.2.1 MusicVAE Generations

Figure 10.2 shows the algorithm for generating blocks from MusicVAE within the Magenta.js library (Roberts, Hawthorne, & Simon, 2018). First, the pre-trained MusicVAE model checkpoint `mel_2bar_small` is initialised. This is because: it is smaller than the other Magenta.js models for faster loading, it outputs single-line melodies which fit wAIve's note grid representation, and it outputs only two bars instead of a longer sequence where only one is needed for a wAIve block.

Next, for each of the workspaces, the user's music is converted from wAIve's grid representation to a *NoteSequence* object provided by the Magenta.js library. This *NoteSequence* is input to MusicVAE's `similar` function. This function encodes the input to MusicVAE's latent space representation, samples a new output from the latent space, and interpolates between them based on the similarity value parsed to the `similar` function. The similarity was set to 0.65 with a temperature of 1.75 as the iterative design found that this created outputs that were similar but not too similar (see pg. 163). The outputs are converted back to wAIve's grid representations, and blocks are added to the workspace.

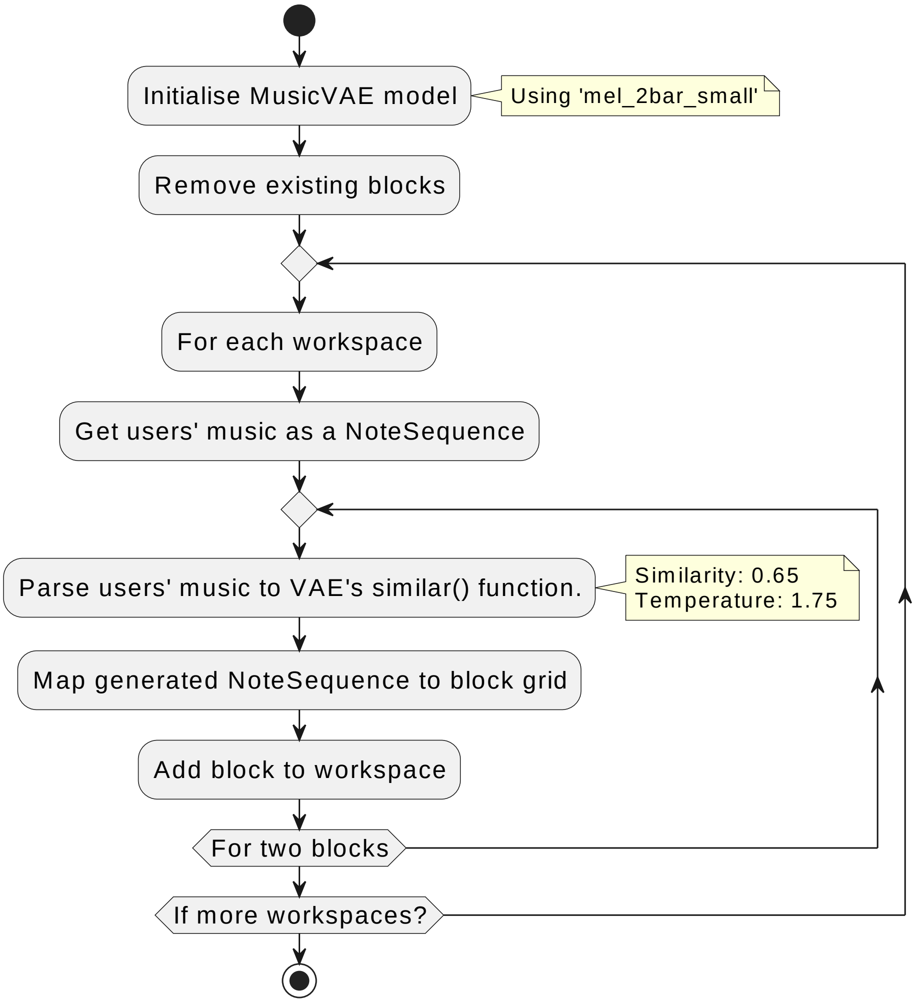


Figure 10.2: Flowchart visualising the sequence for generating music blocks from MusicVAE.

10.2.2 Metric-based Music Clustering

The AI algorithm loaded examples generated from a fine-tuned GPT-2 (Radford et al., 2019) model up to iteration 6 in Chapter 9. These examples were loaded based on how similar musically meaningful metrics were to the user's own music. This followed the approaches in Banar and Colton (2021, 2022), which systematically showed that GPT-2 could generate high quality musical outputs with relation to musically meaningful attributes. This contrasted the available alternatives at the time of the iterative design that produced outputs unrelated to semantically meaningful musical attributes (Bryan-Kinns et al., 2021). For example, whilst FolkRNN had been hosted on a webpage (Ben-Tal et al., 2021), it could only be controlled via a temperature parameter which had little relation to musical values.

A small GPT-2 model (124M parameters) was fine-tuned on all 909 MIDI files from the POP909 dataset (Wang et al., 2020): a dataset of piano arrangements of popular songs. POP909 was selected on the assumption that people find pop music more engaging (Chen & O'Neill, 2020) than classical music (Hadjeres et al., 2017) or folk music datasets (Sturm et al., 2016), which were most commonly used in music AI research at the time of the study[2].

POP909 was pre-processed using Python[3] as follows. First, notes in the C major scale between the MIDI pitches of 60 to 72 and with a rhythmic value equal to an integer multiple of 0.25 were extracted to fit wAIve's note grids. Second, the MIDI data was represented as tokens in the form of "n:xd:y" where x is the note pitch and y is the note duration (Banar & Colton, 2022). Third, this text data was used to fine tune GPT-2 for 3000 steps with a learning rate of 1e-3. 100 generations of tokens with a max length of 256 (temperature = 1.25) were generated. Fourth, the data was converted into wAIve block-sized musical fragments by extracting windows of tokens where the notes summed to a duration of 2.0. This resulted in a database of 581 AI generated musical examples and stored in a JSON format.

The music metrics of pitch count, pitch range, and average pitch interval were calculated for the user's music and the AI database examples. These

[2] Research has since expanded to consider use of more under-represented datasets. For example, see Bryan-Kinns, Fiebrink, et al. (2024).

[3] https://www.python.org/

metrics were chosen to reflect values commonly used in music information research (Banar & Colton, 2022; Yang & Lerch, 2020). The metrics were implemented into wAIve directly as opposed to with an external tool such as JSymbolic (McKay & Fujinaga, 2006) to reduce conversion between MIDI and wAIve's note grid formats. The metrics are defined as follows.

The **pitch count** for each block of music is:

$$\text{Pitch Count} = \sum_{i=1}^{n} \begin{cases} 1 & \text{if } p_i \text{ is unique in the block} \\ 0 & \text{otherwise} \end{cases}$$

where p_i is the i-th pitch in the block, and n is the total number of pitches. This counts 1 for each unique pitch and 0 for duplicates.

The **pitch range** is:

$$\text{Pitch Range} = \max(p_i) - \min(p_i)$$

where $\max(p_i)$ is the highest pitch value and $\min(p_i)$ is the lowest pitch value within the block.

The **average pitch** is:

$$\text{Average Pitch} = \frac{1}{n}\sum_{i=1}^{n} p_i$$

where p_i is the i-th pitch in the block and n is the total number of pitches.

The absolute difference between the users' own music and each database example is then calculated as:

$$\text{Similarity} = |x_{\text{user}} - x_i|$$

where x_{user} is the value of a given metric calculated for the user's music and x_i is the value of the same metric for the i-th entry in the generated dataset. For each metric, the database is sorted by its similarity in ascending order and the top three most similar entries are added as blocks to wAIve's workspace.

10.3 Animations

This section details the flying block, flashing play button, curve, and fading note animations drawn onto wAIve's user interface. Recall that the animations are loaded within wAIve's drawing loop at 60 frames per second.

10.3.1 Flying Block Animations

The flying block animations are executed on an individual AI block chosen at random (see Figure 10.1). When a block starts to fly, a target `X` and `Y` coordinate is randomly chosen for the right-hand side of one of the timelines. Each frame, the `X` and `Y` position of the block is gradually updated based on a percentage, which increases by a small amount (0.015). The increment was based on trial and error from the developer and feedback from users, in the iterative design. When the percentage equals 1.0, the animation stops.

Specifically, the x-position moves in a straight line from:

$$x = \text{startX} + \text{pct} \times (\text{endX} - \text{startX})$$

where `startX` is the initial block position, `pct` is the percentage, and `endX` is the target position. The y-ordinate moves along a curved line using the equation:

$$y = \text{startY} + \text{pct}^4 \times (\text{endY} - \text{startY})$$

where `startY` is the initial block position, `pct` is the percentage, and `endY` is the target position. Twenty-five seconds (based on trial and error and feedback in the iterative design) after the flying block animation, and if the flying block is not clicked, this sequence repeats. However, the bin is selected as the endX and endY to delete the block.

10.3.2 Flashing Play Buttons

Figure 10.3 visualises the algorithm for the button flashing. Each time a play button is pressed, a count is incremented for: the play button for all blocks, the play buttons for the timelines, and the play buttons for individual blocks. When triggered, a sequence of conditions compares the total count of play button presses at the different levels. The flashing is then triggered for the least used button.

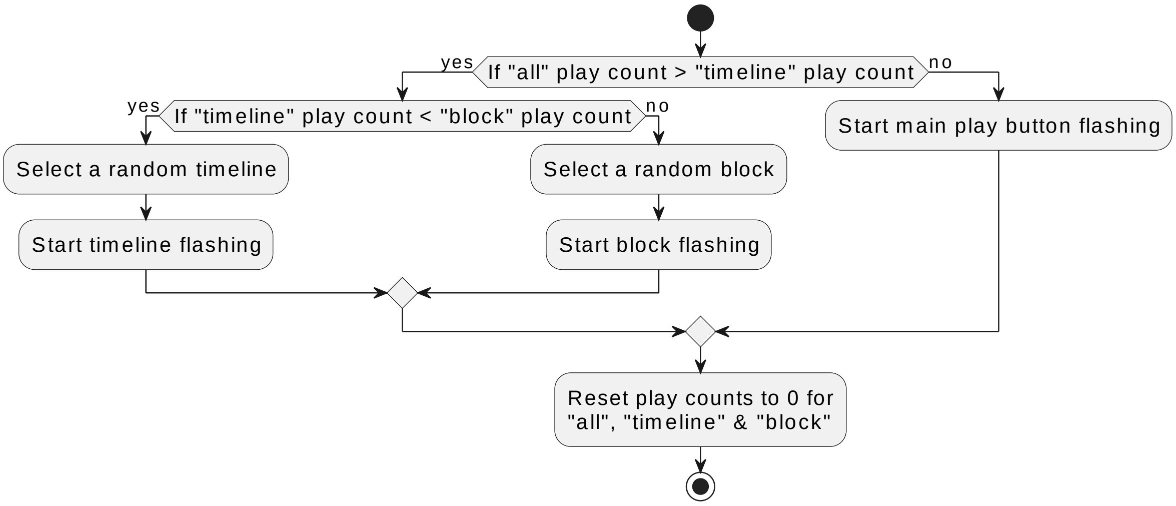

If "all" play count > "timeline" play count
yes
no
If "timeline" play count < "block" play count
yes
no
Start main play button flashing
Select a random timeline
Select a random block
Start timeline flashing
Start block flashing
Reset play counts to 0 for "all", "timeline" & "block"

Figure 10.3: Flowchart of the algorithm used to identify whether blocks should be flashed or not.

The flashing effect is created by modulating the shadow blur of a playback button using a sine function:

$$\text{shadowBlur} = 100 \cdot \sin(\theta)$$

where θ is a floating point variable that increases over time by $\theta \leftarrow \theta + 0.075$. The shadow colour in the final iteration is set to yellow.

10.3.3 Curve Animations

Bézier curves were drawn between the AI blocks and timelines whenever new AI blocks were generated. The left `X` and `Y` of each AI block and the right of the AI music timeline were input to the Bézier curve function. Based on the elapsed time, the transparency value for the curve's colour (alpha) fades linearly after 4 seconds. This process is applied sequentially to blocks for the different timelines. The percentage distance between start and end points for the width and height of the curve equals 0.2, whilst only the bottom timeline curve is set to draw upside-down.

10.3.4 Note Colour Coding and Fading

In iteration 4, a fading animation was used to remove colour-coded note patterns over 25 seconds (see page 160). To achieve this, a `fraction` was calculated as:

$$\text{fraction} = \exp\left(\frac{2}{3} \times \frac{\text{elapsedTime}}{25000\ \mathit{milliseconds}}\right) - 1$$

where elapsedTime is the time that has passed since the animation started. Multiplying by two-thirds slows the transition, whilst subtracting 1 ensures the colour moves towards zero. The exponential function creates a non-linear progression, meaning the colour fade begins slowly but becomes faster over time. The note colour would then be set by interpolating the RGB values of the current colour and the target colour (orange):

$$\text{noteColour} = \text{currentColour} + (\text{targetColour} - \text{currentColour}) \times \text{fraction}$$

10.4 Summary

This chapter described the technical implementation details for wAIve. It showed how the AI generation and animations were implemented using the P5.js and Magenta.js JavaScript libraries. Each feature described was designed to emphasise reflection. This limits the range of possible interactions in wAIve to those focused on reflection, whilst still representing interactions common to AI-based music composition. WAIve thus enables a homogeneous evaluation of users' composition processes. The next chapter presents a user study with wAIve to show how reflection is characterised in its interaction.

Chapter 11

Evaluation of wAIve

The previous chapters introduced wAIve: a novel Artificial Intelligence-based Creativity Support Tool (AI-based CST) for music composition, with reflection support as a central design goal. This chapter evaluates wAIve for reflection and engagement. The evaluation findings use wAIve as a case study tool to identify the interplay between reflection and engagement that is common in people's music composition with AIGC. Interplay here refers to how factors (in this case, reflection and engagement) influence one another. This study is the first to examine the interplay between reflection and engagement in a systematic mixed-methods CST evaluation and the first in the AI-based music composition context. The study confirms that common types of reflection in AI-based music composition align with those identified in existing literature; previously, there was no understanding of their applicability to AI-based music composition in the state-of-the-art.

The chapter is organised as follows. Section 11.1 describes the mixed methods study design. This includes scores from RiCEv2 and participant interviews on their reflection when interacting with wAIve. Section 11.2 presents the study findings with separate analyses for the quantitative (regression analysis) and qualitative (thematic analysis) data. Section 11.3 triangulates these findings to characterise reflection in AI-based music composition and its interplay with engagement.

11.1 Method

This section describes a systematic mixed methods study to evaluate reflection and engagement, using wAIve as a case study. This follows the standard for CST evaluation (Hewett et al., 2005). The study was approved by Queen Mary University of London's ethics committee. Participants gave written consent and were reimbursed with a £20 Amazon voucher.

11.1.1 Participants

Twenty two students studying computer science courses at Queen Mary University of London were recruited. The sample size matches the average for studies with student participants in the CHI conferences before 2016 (Caine, 2016).

Advertisements for the study were presented in lectures on user experience design and physical computing. The degrees reached were computer science, creative computing, design innovation and apprenticeship degrees. While targeting this niche group of students limits broader generalisability, the group was chosen to recruit mostly non-musicians who typically do not experience the benefits of more thoughtful engagement in music composition when casually using a musical CST. These benefits include learning to: question musical intuitions (Kaschub & Smith, 2009), generate ideas from a blank page (Barrett & Hickey, 2003), and develop a musical identity (Barrett & Hickey, 2003; Whittall, 2011). The participants' interest in technology and creativity also meant they would likely enjoy the study. This justifies the sample as a group with some homogeneous characteristics that would benefit from the study's findings.

A description of the participants' demographics and descriptive statistics is shown in Table 11.1. Two pre-test measures were used to capture the participants' expertise.

The **Goldsmiths MSI** (Müllensiefen et al., 2014) assessed the participants' expertise in music. The short scale version was used to avoid fatiguing participants. Figure 11.1 shows the MSI scores. Most participants (18 out of 22) scored below the UK average (86), and the lowest scores were for musical training. This is interpreted as showing that there are mostly non-musicians, with some participants having a keen interest in music (as expected in vol-

Table 11.1: Participants' demographics and descriptive statistics for the wAIve evaluation study.

ID	Age	Gender	Country of Birth	Degree Programme	Study Year	Musician	Instrument
P1	21	Female	UK	Design Innovation	2	Musician	Piano
P2	21	Female	UK	Design Innovation	3	Non-Musician	None
P3	21	Female	Brazil	Creative Computing	2	Non-Musician	Piano
P4	21	Male	India	Computer Science	3	Non-Musician	None
P5	20	Female	US	Creative Computing	2	Non-Musician	Piano
P6	21	Female	UK	Creative Computing	3	Musician	Singing
P7	21	Non-binary	USA	Creative Computing	2	Non-Musician	Drums
P8	19	Male	Italy	Design Innovation	2	Musician	Piano
P9	21	Female	UK	Apprenticeship	3	Non-Musician	None
P10	21	Female	UK	Apprenticeship	3	Non-Musician	None
P11	21	Female	Malaysia	Creative Computing	3	Non-Musician	Voice
P12	21	Female	Spain	Creative Computing	3	Non-Musician	Flute
P13	24	Male	UK	Computer Science	3	Non-Musician	Piano
P14	20	Male	Spain	Computer Science	3	Non-Musician	Piano
P15	20	Male	India	Computer Science	3	Non-Musician	None
P16	22	Male	UK	Creative Computing	2	Musician	Bass
P17	20	Female	South Africa	Design Innovation	2	Non-Musician	Trombone
P18	21	Male	UK	Computer Science	3	Non-Musician	None
P19	22	Male	Pakistan	Computer Science	3	Non-Musician	Piano
P20	21	Female	Ghana	Computer Science	3	Non-Musician	None
P21	21	Female	UK	Computer Science	3	Non-Musician	Piano
P22	20	Female	UK	Apprenticeship	3	Non-Musician	None
	Mean = 20.9	Female (n=13)	UK (n=10)	Computer Science (n=8)	3 (n=15)	Non-Musician (N=18)	Piano (n=8)
	Med = 21	Male (n=8)	Other (n=12)	Creative Computing (n=7)	2 (n=7)	Musician (n=4)	Other (n=6)
	SD = .97	Non-Binary (n=1)		Design Innovation (n=4)			None (n=8)
				Apprenticeship (n=3)			

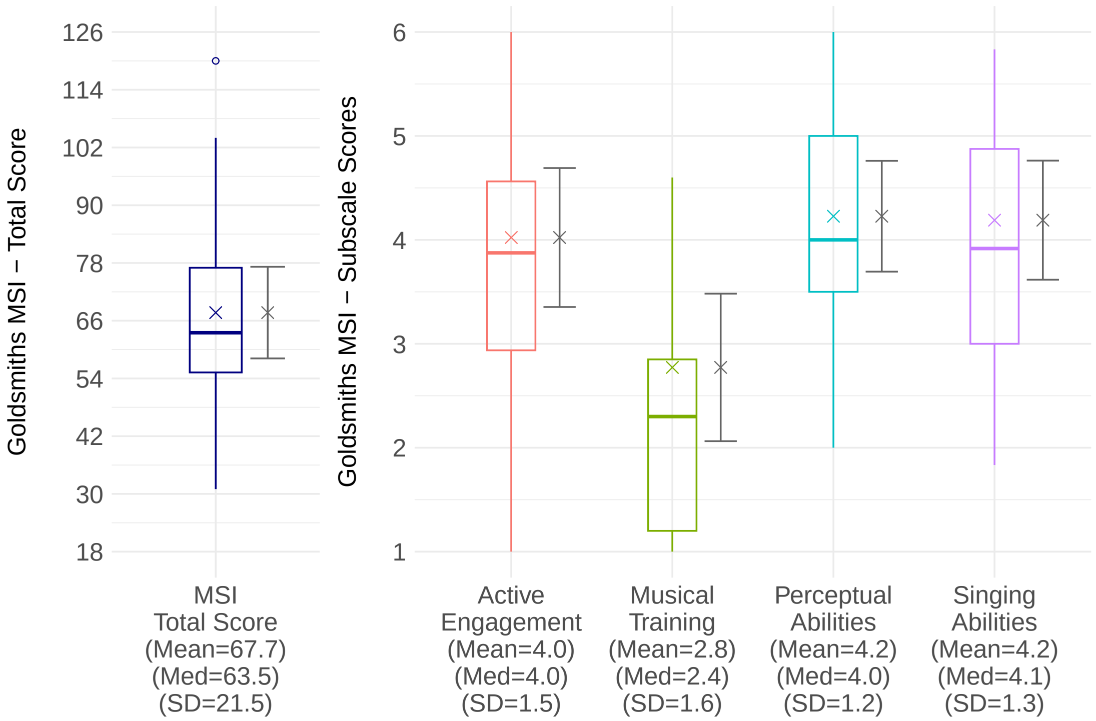


Figure 11.1: Goldsmiths MSI scores in the wAIve evaluation study.

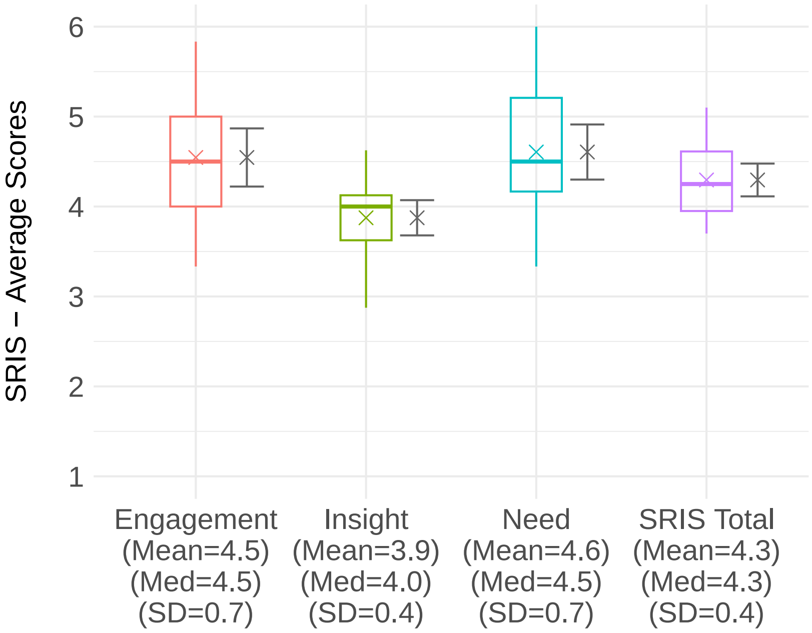


Figure 11.2: SRIS scores in the wAIve evaluation study.

untary studies on music) but little formal training. Musicians are defined as those above the national average, and non-musicians are defined as those below the national average, for later analyses (see Table 11.1)..

The **SRIS** (Grant et al., 2002) assessed participants' capacity for reflection. Figure 11.2 shows the SRIS scores. The average SRIS score (4.3) is similar to the average of the CST users in Chapter 3 (4.5). The subscales show that the participants are generally good at self-reflection but less confident in reaching insights.

11.1.2 Procedure

Each participant met with the researcher individually, in-person, and in a controlled space; no other people were present. They signed the information sheet and consent form. The procedure described below was then followed:

1. **Pre-task Questionnaire:** The participant answers the pre-questionnaire (see Section 11.1.1).

2. **Demonstration:** The participant watches a two minute video demonstrating how to do the following in wAIve: drag blocks, connect and disconnect blocks, move stacks of blocks, add notes, press play at different levels, and delete blocks. No AI features are demonstrated to test how people interact with the AI unprompted. The video is found in the appendix.

3. **Training:** The participants are given up to five minutes to "familiarise themselves" with wAIve. They were instructed that "This is [their] opportunity to ask the researcher any questions before [they] are given the main task.". AI features are turned off at this time.

4. **Task:** The participant is given up to twenty minutes and asked "to freely compose a piece of music with wAIve". This is longer than the task time in previous chapters to provide more opportunity for deeper reflections or feelings of engagement, cf. flow states (Csíkszentmihályi, 1990). The task is purposefully open-ended to explore reflection and engagement in a creative experience with typical characteristics.

   Participants are also told: "You may ask the researcher questions, but they can only offer limited help whilst you compose". The researcher only helps participants if they are stuck on a problem covered in the initial video demo. The AI features of wAIve are turned on at the start of this task. During the task, the participant's computer screen is recorded. The researcher notes moments of interest.

5. **Post-task Questionnaire:** The participant completes the post-task questionnaire (see Section 11.1.3.1).
6. **Interview:** The participant completes a video-cued recall interview (see Section 11.1.3.2).

11.1.3 Data Collection

The study collected quantitative data through questionnaires (Section 11.1.3.1) and qualitative data through a video-cued recall interview (Section 11.1.3.2).

11.1.3.1 Questionnaire Measures

In the post-task questionnaire, the following two questionnaire blocks were collected. For each block, the question order was randomised.

User Engagement Short-Scale: The User Engagement Short-Scale (UEQ) was used to examine aspects of engagement (O'Brien et al., 2018). The UEQ was selected as it focuses on evaluating a technology and does not contain extraneous factors. The short version minimises participant burden and maintains appropriate study duration. Metrics are collected (defined in Section 2.3.2.1) for: focused attention, aesthetic appeal, perceived usability, reward, and a total engagement score.

Reflection in Creative Experience Questionnaire: RiCEv2 was used to examine different aspects of reflection. As described in Table 5.2, metrics are calculated for: reflection-on-current-process (RiCEv2-Cp), reflection-on-self (RiCEv2-Se), reflection-through-experimentation (RiCEv2-Ex) and a total RiCEv2 score.

11.1.3.2 Interview

A video-cued recall interview was conducted (Candy et al., 2006; Candy, 2006) after the post-task questionnaire (see Section 11.1.3.1). Video-cued recall was used to collect participants' reflections on their composing whilst not interrupting their engagement in the moment. The participant's video of their composing was replayed to them, and they were instructed to:

> "Pause at any points where you might have been reflective or thoughtful whilst writing your music".

The instruction was purposefully broad to cover both "thoughtful" and "reflective" moments so as not to restrict data collection.

The researcher guided the VCR interview using a pre-defined set of questions and notes on their own observations. A semi-structured approach was followed to probe unexpected interactions from participants' interactions. The initial questions for the semi-structured interviews are shown in Table 9.2. The questions were more open-ended than for wAIve's design (see Table 11.2) to focus the study on the creative process instead of design features.

Table 11.2: Interview questions used alongside the video-cued recall process in the wAIve evaluation study. Questions marked with \$ were also used for wAIve's design.

Opening Questions
- What are your initial thoughts?$^{\$}$
- Any moments that seemed particularly interesting to you?
VCR Questions
- What were you doing at this moment?
- What was happening at X moment? Why?
- What was your reaction when X occurred?
- — If you felt a certain way, how did you feel?
- Did any of your ideas change? Why?
- How were you deciding what was good and what was bad for your music?
- Were you thinking about other experiences at any moments?
- How did you reflect on your music whilst using the interface?$^{\$}$
- Were there any moments that prompted you to stop to think?$^{\$}$ If so, what?
- Were you consciously thinking these things in-the-moment?
Probes
- Why?
- Would you explain further?$^{\$}$
- Would you give an example?$^{\$}$
- How would you go about explaining this?$^{\$}$
- How did you decide that X was best?
Closing Questions
- Any features that you think are missing?$^{\$}$
- Any general suggestions?$^{\$}$

11.1.4 Data Analysis Method

This section details the analysis methods for the quantitative and qualitative data. The quantitative and qualitative data are analysed separately. This contrasts with a sequential approach, for example, using qualitative findings from interviews to guide a quantitative questionnaire design (Creswell, 2009). Whilst it is not always possible to solve disparities in the findings from analysing data separately, the advantage is that findings are substantiated across the distinct data types (ibid.).

11.1.4.1 Descriptive Statistics

Descriptive statistics are reported for RiCEv2, UEQ, and their subscales to give an overview of the measures. This summarises the sample's scoring for wAIve (Müller et al., 2014). Measures of central tendency (mean and median) and variability (the standard deviation) are provided.

11.1.4.2 Regression Analysis

A linear regression model is used to identify how much a set of variables (engagement) predicts other variables (reflection). Linear regression models consist of the weighted sum of a set of x variables for an output variable y (Kaptein, 2016). Thus, they can identify relationships between variables with only one study condition (Bryan-Kinns & Reed, 2023).

A procedure based on Muller et al. (2020) was followed. First, non-parametric tests are conducted to assess if there are significant differences between confounding variables in the participant pool. Differences between RiCEv2 and UEQ totals are compared for groups within the sample: musicians/non-musicians, degree programs and year of study. When there are two groups, Mann-Whitney U tests (Mann & Whitney, 1947) are conducted, which suit non-parametric data with binary groups. If there are multiple groups, Kruskal-Wallis tests (Kruskal & Wallis, 1952) are performed, which suit non-parametric data but compare three or more groups.

If differences between participants show significant differences, the variable is modelled as a random effect (Kaptein, 2016). Otherwise, a standard linear model is used. Using a training subset of the collected data (14 out of 22 participants), models are created for the RiCE total score and its factors, using the UEQ factors as predictors. Using backwards selection, predictors are removed which improve the model's AIC score (a goodness of fit measure that accounts for the number of parameters in a model (ibid.)). The significant regression equations found are reported. Simplistic models with few variables are strived for, given the small dataset. The quality of the model is assessed using a test set (8 out of 22 participants) and reporting the R^2 value – values towards 1.0 indicate less error.

11.1.4.3 Thematic Analysis

The video-cued recall interview was transcribed using the same process described in 7.1.5. An inductive thematic analysis (Braun & Clarke, 2006; Braun & Clarke, 2019) approach was then performed on the transcripts. The same reflexive thematic analysis approach described in Section 7.1.5 was followed. To recap, the thematic analysis moves between the steps of: generating short descriptive codes for passages in the data, refining codes, organising the codes into themes, and re-applying themes to test their fit. The coding process was applied to identify frequent forms of reflection and engagement that occurred for interactions in wAIve, which are also common to other AI-based musical CSTs.

11.2 Findings

This section reports the findings of the analysis methods described in Section 11.1.4 above.

11.2.1 Descriptive Statistics Findings

Figure 11.3 shows the RiCEv2 scores. RiCEv2-Cp has the highest average. RiCEv2-Se has the lowest average.

Figure 11.4 shows the distribution of scores for the UEQ. Focused attention and reward score the highest average. Reward has two outliers (P6 scoring 5 out of 5; P10 scoring 2.7 out of 5).

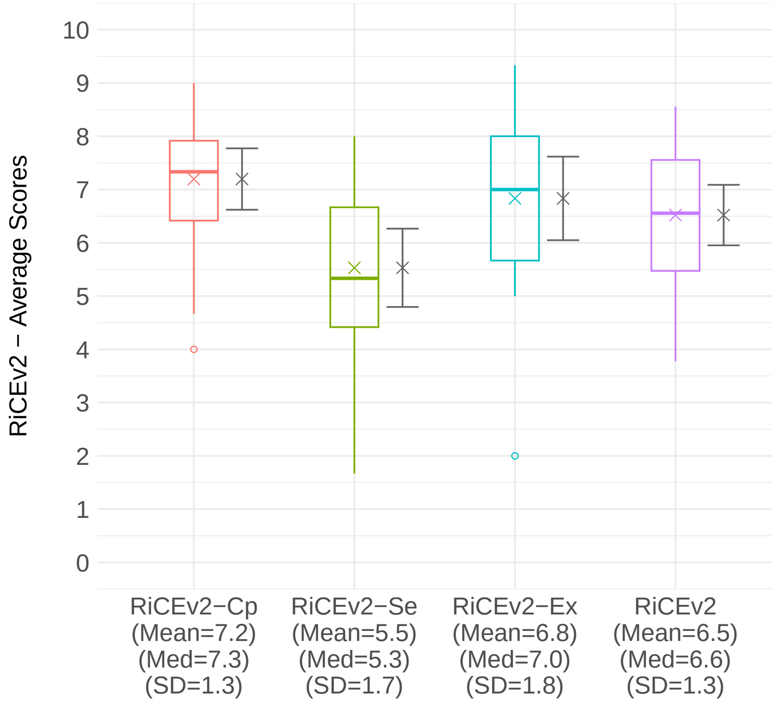


Figure 11.3: RiCEv2 scores in the wAIve evaluation study.

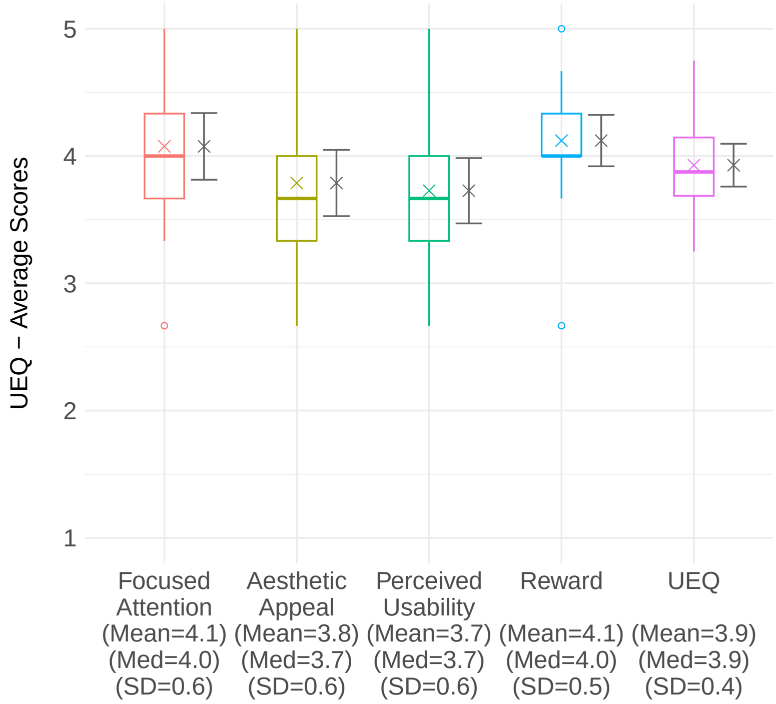


Figure 11.4: UEQ scores in the wAIve evaluation study.

11.2.2 Regression Analysis Findings

The Mann-Whitney U and Kruskal-Wallis tests found no significant differences between RiCEv2 nor engagement for different groups in the study sample (see Table 11.3). This shows no confounding influences across participants' characteristics. Therefore, a simple linear regression (Kaptein, 2016) is performed with no random effects.

Table 11.3: Comparisons between the UEQ total score metric and RiCEv2 total score metric in the wAIve evaluation study.

Group	Test	Measure
Musicians	Mann-Whitney U	**UEQ:** W = 44.5, p = .47 **RiCEv2:** W = 56, p = .09
Year of Study	Mann-Whitney U	**UEQ:** W = 53, p = .97 **RiCEv2:** W = 70, p = .22
Degree Programme	Kruskal-Wallis	**UEQ:** χ^2 = 3.04, df = 3, p = .39 **RiCEv2:** χ^2 = 6.90, df = 3, p = .08

Four significant regression models are identified, reported in Table 11.4, and visualised in Figures 11.5 to 11.8. The best performing model overall is model 1 – participants who reflected more found the experience to be most rewarding. The best model in the training set is for RiCEv2-Se (model 4), which has a significant linear relationship with reward. However, the R^2 is small in the test set.

Table 11.4: Significant regression models for the wAIve evaluation study. Aspects of engagement predict aspects of reflection.

		Training Set			Test Set
Model	**Predictor**	**Equation**	**Adjusted R2**	**F-statistic**	**R2**
1	RiCEv2	-3.99 + (1.11)Focused_Attention + (1.41)Reward*	0.57	F(2,11) = 9.55 p < .001*	0.50
2	RiCEv2-Ex	-4.06 + (2.70)Aesthetic_Appeal*	0.39	F(1,12) = 9.32 p = .010*	0.33
3	RiCEv2-Cp	-0.82 + (2.10)Focused_Attention* + (1.07)Reward - (1.35)Aesthetic_Appeal	0.45	F(3,10) = 4.54 p = .030*	0.16
4	RiCEv2-Se	-7.09 + (1.94)Reward*	0.63	F(1,12) = 23.22 p < .001*	0.31

* = p < 0.05

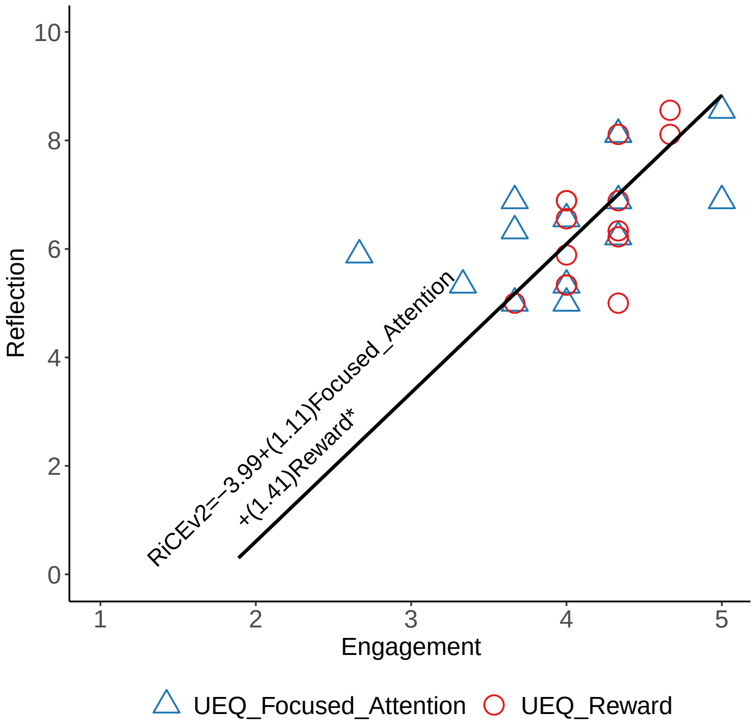


Figure 11.5: Model 1 visualisation of RiCEv2 ∼ UEQ factors.

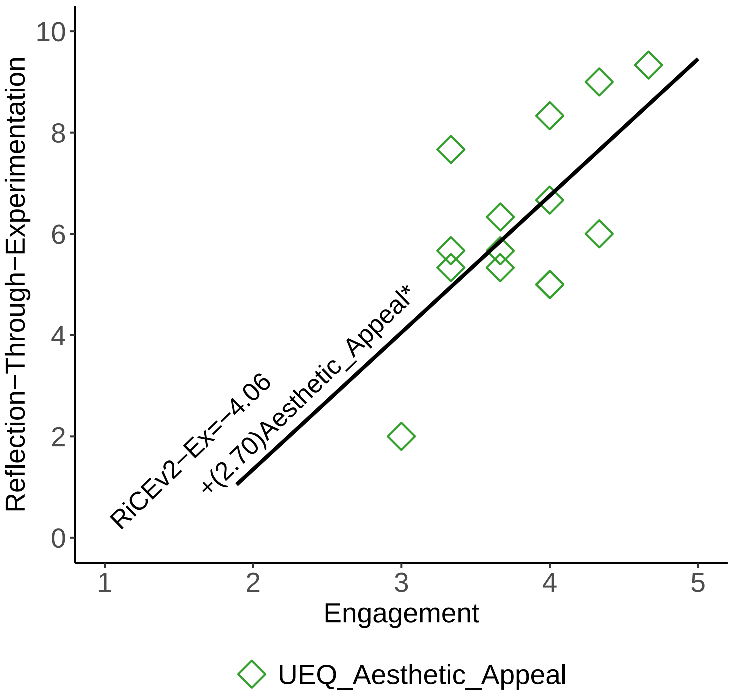


Figure 11.6: Model 2 visualisation of RiCEv2-Ex ∼ UEQ aesthetic.

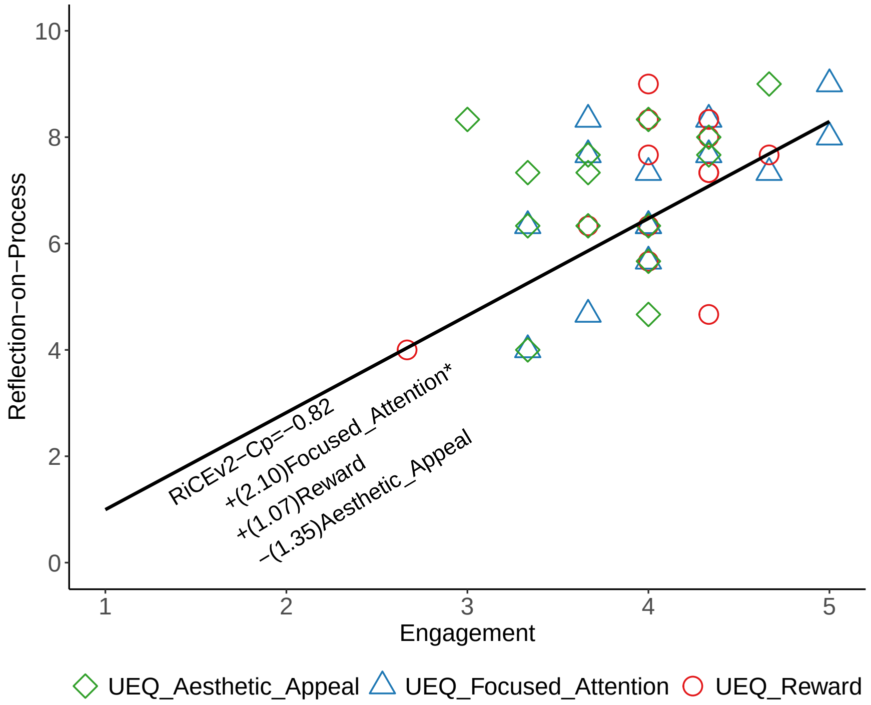


Figure 11.7: Model 3 visualisation of RiCEv2-Cp ∼ UEQ factors.

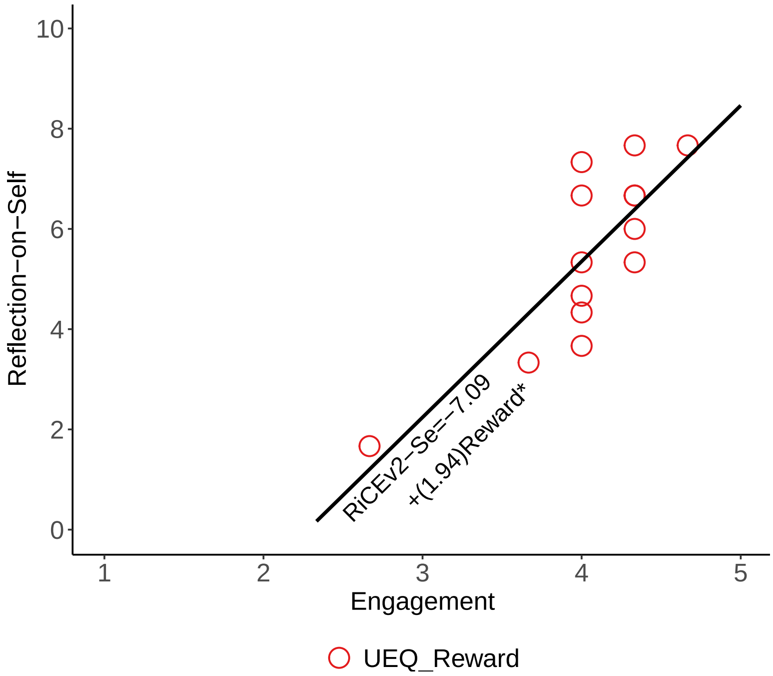


Figure 11.8: Model 4 visualisation of RiCEv2-Se ∼ UEQ reward.

11.2.3 Thematic Analysis Findings

Seven themes were generated from the thematic analysis of the video-cued recall interview transcripts. These are described below.

11.2.3.1 Theme 1: Participants' motivation informed their reflection

Participants had different motivations when approaching the open-ended task of composing a piece of music. These motivations influenced the aspects of wAIve that participants reflected upon. The main motivations were to create quality music (herein named *music-first*) or to have lots of fun (herein named *fun-first*).

The majority of participants (P1-P5, P7-P16, P18-P21) had a music-first motivation. They wanted to create music that sounded "good" (P19) or "nice" (P7). Some participants directly outlined this motivation. For example, P15 said "I wanted to create something on my own[...] something which is fun to listen to". P11 said "I kind of want to have[...] a result out there". Music-first participants reflected on the musical qualities of their composition to improve its quality. For example, P2 reflected on note patterns that would sound "nice" ("it's very nice to listen to notes going up and down"). P18 reflected on different instruments that "made [their music] sound a whole lot better".

Fun-first participants (P6, P22, P17) prioritised their enjoyment over the quality of the music. For example, P6 said that "any task where I was enjoying myself, I was going to be happy. I am happy with the outcome regardless of its quality". P20 said "I don't think I would really care much about who would listen to" their music. P17 said "I think it sounded good, but I wasn't too concerned[...] because it was fun". Fun-first participants performed interactions to experience more autotelic engagement rather than emphasising reflection.

Participants' motivations also shifted from fun-first (when initially interacting with wAIve) to music-first. P22 showed playful and creative behaviour, adding note patterns matching the letters of their name because "I wanted to know what my name sounds like" (P22). After an AI suggestion flew in, P22 "felt weird[...] because I was like, well, this is not what I'm supposed to do". They considered "is this what they [the AI] want?' Am I doing it wrong?". P15 also started by "*playing* [emphasis added] with and experimenting with all the tools", until "the last 10 minutes[...] when I actually tried to do something to make it all sync".

11.2.3.2 Theme 2: Reflecting on animations sparked music learning and reflection-through-experimentation

There were several moments where wAIve showed interactions to the participants that they "never would have done" (P1) otherwise. This led to moments of learning amongst participants. For example, participants observed and then recreated the flying AI blocks' animations. P21 saw the AI blocks "flying and overlapping", and was motivated to test this, realising "I can do that too". P17 observed the AI animation and realised that the colour of blocks did not restrict blocks to only the same coloured timeline. This sparked their reflection-through-experimentation to "switch things out" and move AI blocks between timelines.

Many participants learnt to move blocks to the start of a block stack by following the animated curves. This is evidenced by P12, who "did imagine it [the curves] means this [block] could go here". P22 "didn't understand the purpose of why [the lines] were there or what they wanted me to do". However, later they were motivated to experiment with moving a block to the end of a curve to "see if I could do it". This contrasts with the intention of the feature from the iterative design to show relationships between the AI music and users' music.

11.2.3.3 Theme 3: Relationship between focused attention and reflection on AI

Participants experienced moments of focused attention, a factor of engagement (O'Brien et al., 2018). For example, P1 was "hyper-fixated", whilst others (P7-P9, P11, P12, P14, P17-P21) described moments of "focus" when working on and listening to note patterns. In particular, moments of focused attention occurred when people listened to their music from different perspectives. For example, P3, P5, P6, P8, P9, P19 and P21 focused on hearing the music as a whole, to test that "at the end[...] it has to make sense all together" (P9).

In some cases, participants were too immersed in moments of focused attention to notice AI animations. For example, P11 said they were “very focused on this part [adding notes, and thus] I didn’t see anything that happened elsewhere”. Others purposefully clicked on the AI to put it out of their mind and stay focused on the task. For example, P7 described that “the [newly suggested AI] segments [are] great, let me go back to my part”. P11 said “I’ll only need that [AI generated block] anyway, so let me put it here”.

Participants also described the AI block animations as “brief” (P15), not “overwhelming” (P1) or “not annoying, just[...] there” (P11). However, there were several moments where the AI broke participants’ attention whilst writing blocks of music. P2 said they were “in my little creative moment, and it [the AI] came and said hi out of nowhere”. P8 said “I was trying to focus on one part[...] and things [the AI] just came[...]. I don’t want that at all. Go away.”

The AI features also helped participants at moments where their focused attention was already broken, such as when “feeling uninspired” (P12), “stuck” (P18) or when “running out of” ideas (P19). P8 and P6 found the AI was helpful during these moments because it sparked their reflection on future directions (reflection-on-process) for their music. For example, P8 said the AI “gave me[...] a few directions where I can take [the music]’.

11.2.3.4 Theme 4: AIGC reduced reflection on unfamiliar instruments

The participants’ previous experience playing musical instruments reduced their reflection on wAIve’s instruments. For example, P7 focused on the instruments they were most familiar with, relying on the AI to write parts for unfamiliar instruments. They noted that they “did focus a lot on the drums just cause that’s my speciality” and would “let the AI do more on the bass and the guitar[...] actively resisting [the AI] a little bit on the drum”. P17 noted that they “were most comfortable with[... the guitar] and[... the drums] because I started learning [the drums]” and also focused their experimenting on their most familiar instruments.

11.2.3.5 Theme 5: Participants’ expectations of AI changed their willingness to reflect

The participants’ expectations for wAIve’s AI informed whether they would reflect on it. P2, P10, P13, P17, P18 and P19 assumed that the AI was better than they were at writing music. P2 said the AI generated music is "probably better than what [they're] making". P10 assumed that the AI is "gonna tell [them] the right suggestion".

This led to moments where participants used the animated AI blocks without reflecting on how they sound. They assumed that the material was immediately usable. For example, P10 said they would "roll with" different AI generated blocks as they flew in. P20 said the AI animations were "cool because [they] don't have to *think* [emphasis added] about exactly what [they're] doing next". P5 felt the AI should be left alone: they "just figured that[...] if [the AI] wants to do that[...] I'm like you do you". Similarly, P21 said, "if [the AI] wants to be there, it can".

11.2.3.6 Theme 6: Curiosity sparks reflection on AI

P6, P14, P16 and P17 were curious about wAIve's AI, sparking their motivation to reflect on how it works. P14 was motivated to reflect on the AI as they "thought it would generate something based on the first two [blocks... so] was curious what it would do". This demonstrates their curiosity about the direction in which the AI will take their music. P16 "was just curious to see how I could make [the AI] fit into what I made" – reflecting on how the AI's behaviour linked to their own.

Other participants hinted at being curious about the AI generated blocks (P2, P3, P4, P6, P8, P10, P14, P17); however, they did not reflect on how the AI worked per se. Instead, they were motivated to reflect on "what they sound like" (P2). P10 reflected on the AI's timing ("let's test what [the AI] does and see what time they would produce") and how the block fit in with "the overall thing and [how] it sounded". They were motivated by a curiosity to hear how the blocks sounded in the broader context.

11.2.3.7 Theme 7: Similar blocks motivated reflection

Several participants (P3, P6, P8, P9, P11, P17, P22) searched for AI blocks visually similar to their own blocks. P11 searched for AI blocks that they thought "would sound good because it looks the same". P18 tested AI suggestions when they were "similar to what I originally had". However, slight differences in the music were also desired. For example, P22 "didn't want to mimic [the AI] exactly". P9 "liked the fact that [an AI suggestion] was a little bit different". There was also a preference for AIGC, which has many notes. P3 said they didn't like the AI blocks because "it's very sparse, it doesn't have much". P6 found there were AI blocks that they "[weren't] so interested in because there wasn't anything on them". Here, the participants used their observations of how the blocks looked in lieu of listening to them.

11.3 Discussion

Table 11.5: Summary of Chapter 11's main findings.

Main Finding	Location
Reflection is characterised in AI-based music composition by people's perception of an AI. Reflection occurred when people perceived the AI as less skilled than themselves.	§11.2.3.5
Reflection is characterised in AI-based music composition as occurring when users discover new ways to interact with the CST, based on the AI's animations or outputs.	§11.2.3.2
Reflection is characterised in AI-based music composition as enabled by people's motivation and curiosity. Participants with a music-first motivation and a curiosity about how the AI worked were more likely to reflect on AIGC.	§11.2.3.1 §11.2.3.6
The interplay between reflection and engagement is that moments of reflection occur alongside moments of focused attention.	§11.1.4.2 Figure 11.5
The interplay between reflection and engagement is that self-reflection occurs alongside feelings of a rewarding user experience.	§11.1.4.2 Figure 11.8
The interplay between reflection and engagement is that AIGC is used without reflection in moments of hyperfocus.	§11.2.3.3

This chapter identified the interplay between reflection and engagement in computer science students' music composition with AIGC. WAIve was used as a case study tool to identify reflection patterns common to AI-based music composition. The main findings are summarised in Table 11.5. Quantitative analyses of reflection and engagement measures found regression models that show interplay between their factors. A thematic analysis of video-cued recall interviews also characterised how the AIGC influenced reflection.

The findings are triangulated to confirm that patterns of reflection from literature also occur in AI-based music composition below. Section 11.3.1

focuses on the interplay of reflection and engagement. Section 11.3.2 focuses on the influence of AIGC. Through discussion of the study findings, characterisations of interplay between reflection and engagement are introduced (see Figure 11.9 and Figure 11.10). Section 11.3.3 describes the study limitations.

11.3.1 Interplay between Reflection and Engagement

Figure 11.9 visualises the interplay between reflection and different types of engagement based on the study findings. The aspects of the model are discussed in the following subsections.

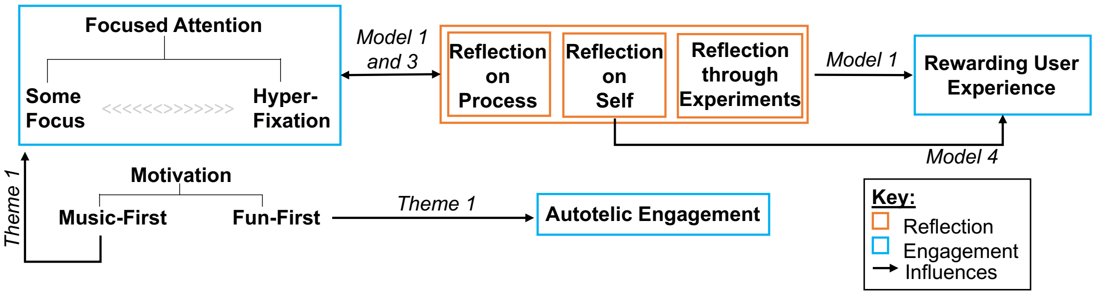


Figure 11.9: Model of interplay between reflection and engagement. Types of engagement are in blue. Types of reflection are in orange.

11.3.1.1 Focused Attention and Reflection

There is interplay between moments of reflection and focused attention (a type of engagement). Focused attention was significant in regression model 3; this shows a relationship between focused attention and reflection-on-process. However, although model 3 is the second most accurate model on the training set data ($R^2 = 0.45$), it performed poorly on the test set data ($R^2 = 0.16$). This shows limited generalisability to similar participants. The pattern from the model visualisation in Figure 2.2 is also unclear. It thus cannot be claimed that there is a link between focused attention and reflection-on-process in all cases.

Regression model 1 shows more substantial evidence of overlap between focused attention and reflection on the whole (measured by RiCEv2). The model shows RiCEv2 as the sum of focused attention and reward measures from the UEQ, and attains the highest accuracy scores in both the training

($R^2 = 0.57$) and test set data ($R^2 = 0.50$). Regression model 1's scatter plot (see Figure 11.5) further shows an observable pattern between focused attention and RiCEv2. However, focused attention is not a significant predictor in model 1, limiting the evidence that this relationship was not by chance. Nonetheless, the overlap between focused attention and reflection corroborates literature characterising flow states as periods of intense concentration (Nakamura & Csíkszentmihályi, 2009). This confirms that reflections on the creative process, personal meaning and through experimentation – evidenced by the questions of RiCEv2 – occur alongside focused states in AI music composition contexts.

11.3.1.2 Reflection and Rewarding User Experiences

There is evidence of a relationship between reflection and whether participants had a rewarding and worthwhile experience (indicated by the UEQ factor of reward). Model 1 (with the highest accuracy scores) found reward to significantly predict the total RiCEv2 score. The scatter plot in Figure 11.5 also shows a link between reflection and reward. This is corroborated by related work. For example, flow theory describes that working at the peak of one's ability is a positive and rewarding experience (Csíkszentmihályi, 1990; Nakamura & Csíkszentmihályi, 2009; Seligman & Csíkszentmihályi, 2014) Costello and Edmonds's (2007) pleasure of difficulty characterises that enjoyment comes from having developed skills to complete a task. Music composition literature has also suggested that the challenges of music composition, such as overcoming a blank page (Barrett & Hickey, 2003) or questioning musical intuitions (Kaschub & Smith, 2009), lead to beneficial and rewarding experiences. This confirms that interplay between reflection and the engagement factor of reward occurs in the AI-based music composition context.

Model 4 of the linear regression analysis also shows that higher levels of *self*-reflection (indicated by RiCEv2-Se) occur alongside more rewarding experiences. This corroborates Hubbard et al. (2021) and Li et al. (2023), who both found that children who reflected on personally meaningful experiences with AI had rewarding learning experiences. It also supports the discussion in Chapter 3 that considering the perspective of others is less important than self-reflection in CST interaction (see Section 3.4.1). However, RiCEv2-Se had the lowest average score overall (see Figure 11.3) compared with the other factors. This confirms that whilst important and rewarding, self-reflection occurs infrequently in AI-based music composition.

However, model 4 is skewed by the participant with the lowest score for reward in the training set (see the leftmost circle in Figure 11.8). These scores were from P10, who also gave the lowest scores for their musical sophistication (MSI = 31). Whilst a direct causation cannot be claimed, their low scores for whether the experience was rewarding occur alongside a lack of interest in the study's musical task.

11.3.1.3 Motivation, Focused Attention and Autotelic Engagement

The study found a relationship between participants' motivations and whether they experience moments of focused attention, or more autotelic types of engagement (Theme 1). For example, some participants approached the task from a fun-first perspective, experiencing moments of autotelic or passive engagement by casually exploring wAIve cf. Compton and Mateas (2015). For example, P6 said "any task where I was enjoying myself I was going to be happy". This confirms other studies on music interaction, which show the effect of task motivation (Wu & Bryan-Kinns, 2019).

In contrast, the participants who focused on creating a quality piece of work (music-first) experienced more moments of focused attention, such as being attentive to notes in their music. For the music-first participants, the level of focus also varied in intensity: from moments of some focus up to moments of hyper-fixation (for example, see P3 in Theme 3). This confirms theories of engagement which consist of multiple levels (Doherty & Doherty, 2018; Edmonds et al., 2006). The intensity of participants' focused attention also changed depending on whether participants were open to interruptions from AIGC in wAIve, discussed further below.

11.3.2 Reflection and AI Generated Content

In addition to interplay between reflection and engagement, the study findings confirm that AIGC influences reflection. Figure 11.10 extends the visualisation in the previous section (Figure 11.9) to show how AIGC influences reflection in AI-based music composition.

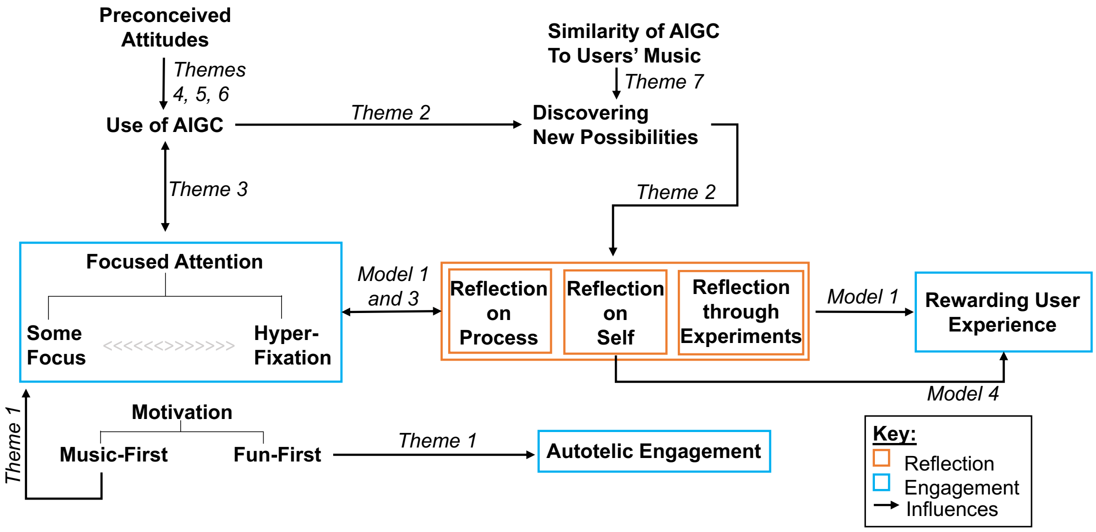


Figure 11.10: Model extending Figure 11.9 to show the influence of AIGC on reflection and engagement. Types of engagement are in blue. Types of reflection are in orange.

11.3.2.1 Use of AIGC and Focused Attention

There is complicated interplay between focused attention and moments of reflection, relating to how participants used the flying block animations. The flying block animations give insight into typical interactions in AI-based CSTs by mimicking the common turn-taking interaction style (Rezwana & Maher, 2022). The interplay between interruptions and focused attention depended on participants' level of focus (as identified in Theme 3) and their motivation (see Theme 1).

When focused but not hyper-fixated, participants would dismiss AI animations to put them out of their minds or save the block for later. For example, P7 said that an AI generated block was "great" and then "let me go back to my part". However, AI animations still posed an unwelcome interruption when participants were hyper-fixated on their music, disrupting their sustained engagement and flow states (Csíkszentmihályi, 1990). Only outside of hyper-fixated states, such as during moments when feeling "uninspired" (P12) or "stuck" (P18), were suggestions from AI welcomed.

Kahneman's (2011) theory of fast and slow thinking provides an interpretive lens for analysing how AIGC supported or distracted people from reflection during their focused attention. The AI animations triggered moments where quick decision making was required, for example, whether to dismiss a block or to keep it (see Theme 3). The time frame for deciding to use an AI block was short; it thus required fast thinking. This is supported by Theme 7 which shows that participants identified similarities between the AI generated music and their own without listening to the block; they used aspects of the visual representation as a proxy for their listening. This demonstrates that they were replacing a more difficult contemplation on the musical quality of a block with an easier one, cf. fast thinking (Kahneman, 2011).

11.3.2.2 Preconceived Attitudes and Use of AIGC

Different preconceived attitudes influenced participants' motivation to reflect on AIGC. Theme 6 shows that participants needed to be motivated by a curiosity about AI to reflect on wAIve's blocks. This confirms Fleck and Fitzpatrick's (2010) and Slovák et al.'s (2017) findings that showing people more data to reflect upon (such as the AI generated blocks) does not necessarily spark their reflection for the AI-based music composition context.

Participants had different attitudes towards AI and preconceived expectations of its role in the creative process. For example, many viewed the AI as a musical expert and trusted its decisions without critique (see Theme 5). For example, participants would "roll with" (P10) AIGC and not consider next steps. These participants made decisions through intuitive, fast thinking instead of slower, more contemplative thinking, where critique more frequently occurs. This confirms that Kahneman's (2011) fast thinking occurs in AI-based music composition. This also confirms Reicherts et al.'s (2022) finding that people who perceived an AI as an assistant let it "do the thinking" on their behalf for AI-based music composition.

Furthermore, the finding that perception of an AI hinders reflection is supported by Theme 4. Theme 4 shows that some participants preferred using AIGC for instruments they were least familiar with. Their lack of skills with an instrument made them believe the AI would act on their behalf. This discourages people from building skills they are unfamiliar with or building the necessary expertise that leads to flow states (Csíkszentmihályi, 1990).

11.3.2.3 Use of AIGC and Discovery of New Possibilities

The participants observed the AI animations to discover new methods of interaction. The AI prompted their reflection on newly found opportunities (see Theme 2). In reflecting on animations to discover new ideas, they were exposed to a broader range of musical possibilities than they had considered possible with wAIve. This corroborates research on CSTs, which allowed people to replay their creative process to reflect on and learn techniques from other creative practitioners, such as 'Watch me write' (Carrera & Lee, 2022) and Spin (Tseng & Resnick, 2016). This also shows an example of Costello and Edmonds's (2007) pleasure of discovery in the AI-music composition context.

11.3.3 Limitations

There are clear variations in the data collected, such as in the musical expertise and courses studied by the participants. Whilst effort was made to control for variation and no significant differences were found between subgroups (see Table 11.3), the qualitative findings show clear differences in areas such as task motivation (cf. Theme 1) which were not quantified. The Queen Mary University of London students and their specialist degrees are a niche cohort of test subjects. The results thus show limited generalisability to the broader population. However, the sample's homogeneity supported the quantitative analysis approaches and show less variation than the artist-researchers in Chapter 8.

The participants were also recruited from classes taught by the thesis author. The participants have a motivation to appease the author and thus could have given overly positive feedback. A novelty effect is also acknowledged as generative AI was emerging into the mainstream at the time of study. Indeed, Theme 6 shows that participants' curiosity about AI sparked their reflection.

The study focused on a contrived musical task captured in a controlled lab setting. Whilst mimicking the open-ended way a non-musician is expected to interact with a CST casually, this poses limited ecological validity. This contrasts with the more ecologically valid approach of observing musicians in their typical places of happening from Chapters 7 and 8.

However, the study is the first to systematically show interplay between reflection and engagement in a typical CST mixed-methods user study. The study findings confirm that the patterns of reflection identified corroborate related literature. Generalisability and rigour are also added to the findings through triangulation (Bryan-Kinns & Reed, 2023). WAIve, as a case study tool, has characteristics common to AI-based CSTs. For example, the AI animations mirror the turn-taking interaction style of many AI-based CSTs (Rezwana & Maher, 2022). Its DAW layout is also common across AI-based music composition tools (Hunt et al., 2020; Tchemeube et al., 2022). The advance on the state-of-the-art is that the observed patterns of reflection are supported by other literature and demonstrated as applicable to AI-based music composition; previously, there was no systematic evidence characterising reflection in this context.

The decision was made to conduct an open ended study, instead of following a comparative approach. WAIve represented typical interaction patterns in AI-based CSTs and confirmed that related patterns from literature were present in the thesis' case study context. This best fits with the thesis's research questions (Section 1.3) on characterising reflection and identifying interplay because it enabled a range of reflection patterns to be observed; a comparative study would be limited to observing singular effects. Indeed, deciding which elements of wAIve to isolate was unclear following its design in Chapter 9. For example, if the effect of the AI were isolated, it would still be unclear whether its animations, style of music generation, or different visual elements affected reflection. Testing wAIve with and without AI could thus only demonstrate that some aspect of the AI design affects reflection more or less than composing with a step-sequencer. This would give little insight into how reflection is characterised in the multifaceted AI-based music composition context. The advance on the state-of-the-art in this study, which confirms that specific patterns of reflection are present in AI-based music composition, could indicate how to design a controlled study. For example, comparison between different timings for an AI interruption (see Section 11.2.3.3) could show differences in engagement and reflection.

This study investigated how people interact with wAIve for the first time and for a limited time. There are temporal aspects to reflection and engagement (Bilda et al., 2008; Boud et al., 1985; Kolb, 1984), as identified in Chapter 8. Longitudinal studies would need more sophisticated software with a wide range of features to examine engagement without users becoming bored cf. flow theory (Csíkszentmihályi, 1990). However, wAIve was intentionally designed with limited functionality to investigate engagement in musical interaction within one controlled session. This ensured that all participants began with a consistent baseline of knowledge and familiarity.

11.4 Conclusion

This chapter evaluated reflection and engagement in AI-based music composition, using wAIve as a case study tool. Using mixed-methods and applying the RiCEv2 questionnaire, computer science students' interaction with wAIve was examined. The findings identified interplay between reflection and engagement in wAIve's interaction. Indeed, models of interplay between reflection and engagement and the influence of AIGC on reflection were presented to characterise reflection in this context. The findings emphasise the correlation between focused attention and reflection, the importance of self-reflection, and that AIGC best supports reflection when users are motivated and can learn from the AI interaction. The following chapter compares the RiCEv2 measures with the other tools in this thesis, consolidating the findings across studies.

Chapter 12

Comparison of RiCEv2 Results

The previous chapter evaluated wAIve to identify interplay between reflection and engagement in AI-based music composition. This chapter analyses the RiCEv2 measures across the thesis. This includes the comparison of wAIve's RiCEv2 results with the RiCEv2 results from other tools. From the comparisons, characterisations of reflection that are common in different tools, including AI-based music composition tools, are identified.

The chapter is organised as follows. Section 12.1 describes the differences between participants and study settings across the chapters in the thesis. Section 12.2 through to 12.5 then analyses the total RiCEv2 score, and RiCEv2 factor scores, in turn.

12.1 Participants and Study Settings

There are similarities in the data collection and analysis methods used across the thesis's studies. Each study used a similar open-ended task, such as to "freely compose" a piece of music, as is characteristic of creative user experiences (Kerne et al., 2013). The studies differ in the selection of participants, the CSTs tested, and their settings. To further understand the differences between participants, the Goldsmiths MSI scores (Müllensiefen et al., 2014) and SRIS scores (Grant et al., 2002) are inspected across studies. This compares the musical expertise and capacity for reflection for each subset of participants.

Figure 12.1 visualises the MSI scores from Part II and Part III of the thesis (Part I focused on CST interaction more broadly and was not music focused). The participants from the design and evaluation of wAIve are similar in their musical sophistication. The artist-researcher participants in Chapter 8 score higher for musical sophistication and show less variation. An exception is the two AI and Music PhD students who supported the design of wAIve in Chapter 9. They are closer to the Chapter 8 participants in their background. The sub-scales of the MSI show that the main difference between the participants is in their formal musical training.

Figure 12.2 shows the SRIS scores across all studies. The participants for the iterative design of wAIve and the artist-researchers are more naturally reflective. This reflects the recruitment strategies: the artist-researchers were PhD students, and the iterative design of wAIve sought participants comfortable with offering critiques. The wAIve evaluation participants are the least naturally reflective.

12.2 Reflection

With an understanding of the differences between the study participants as context, RiCEv2 scores are compared for the CSTs evaluated in the thesis. RiCEv2 offers scores for reflection-on-process (RiCEv2-Cp), reflection-on-self (RiCEv2-Se) and reflection-through-experimentation (RiCEv2-Ex). Developed in Part I of the thesis, these types of reflection corroborate related work such as Candy's (2019). For example, reflection-through-experimentation is similar to reflection-in-the-making-process in that they both refer to decision making during interaction. Reflection-on-process is also similar to reflection-for-action in that it refers to points where people decide how to progress their making going forward. However, reflection-for-action occurs at the start of the process instead of during interaction, contrasting reflection-on-process. The main difference is that RiCEv2 operationalises reflection as a construct that is measurable by questionnaire items; Candy's (2019) reflection types were derived from qualitative interviews and do not produce measures directly from users, nor can they be numerically inspected.

To enable comparison, the RiCEv2 metrics are averaged from each user study. The artist-researcher's scores show the average across four compo-

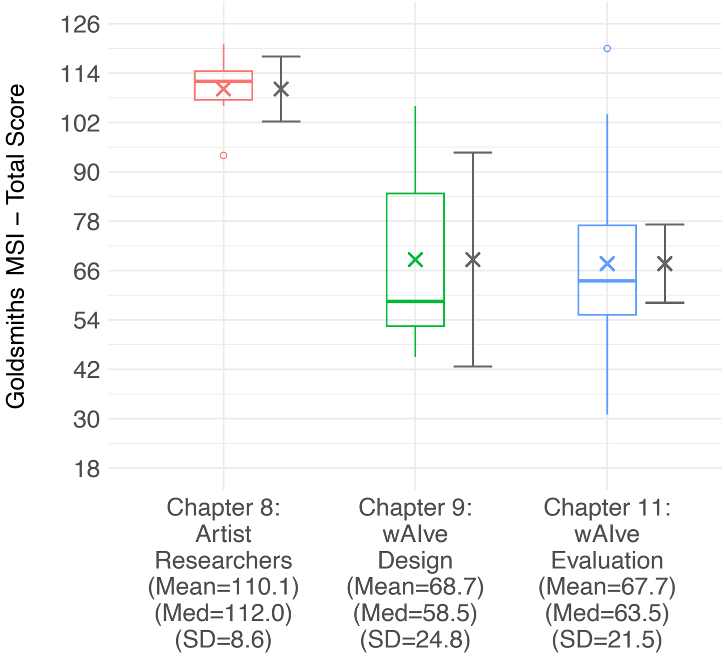


Figure 12.1: Goldsmiths MSI scores for CSTs across chapters.

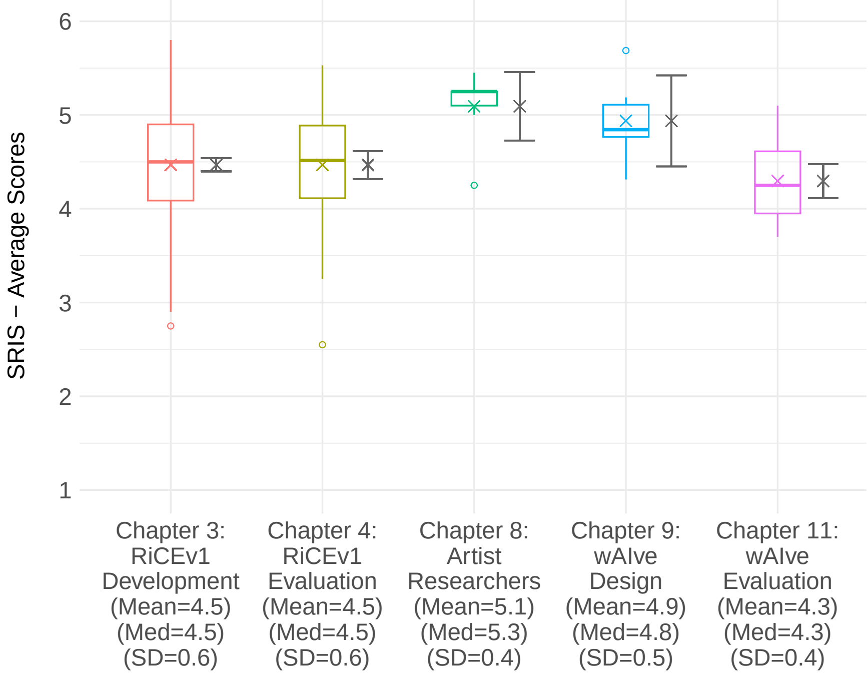


Figure 12.2: SRIS scores for CSTs across chapters.

sition sessions. The other scores represent a single post-hoc assessment of interaction with a CST. The CSTs selected from Chapter 3 are MS Word, Photoshop, Visual Studio and DAWs. This balances selecting CSTs: with the largest proportion of participants, representing a range of different creative practices, and that can draw comparisons between music and AI-based music tools. The scores from story-sentiment-visualiser and sound-sketcher in Chapter 4 are not included as only RiCEv1 results can be calculated from the study's data.

Figure 12.3 shows the total RiCEv2 scores for the CSTs. The means are close together (approximately 6.8) regardless of participants' backgrounds or study settings. A Kruskal-Wallis test (Kruskal & Wallis, 1952) shows no significant difference between RiCEv2 scores across the CSTs (χ^2 = 5.27, df = 6, p = .51).

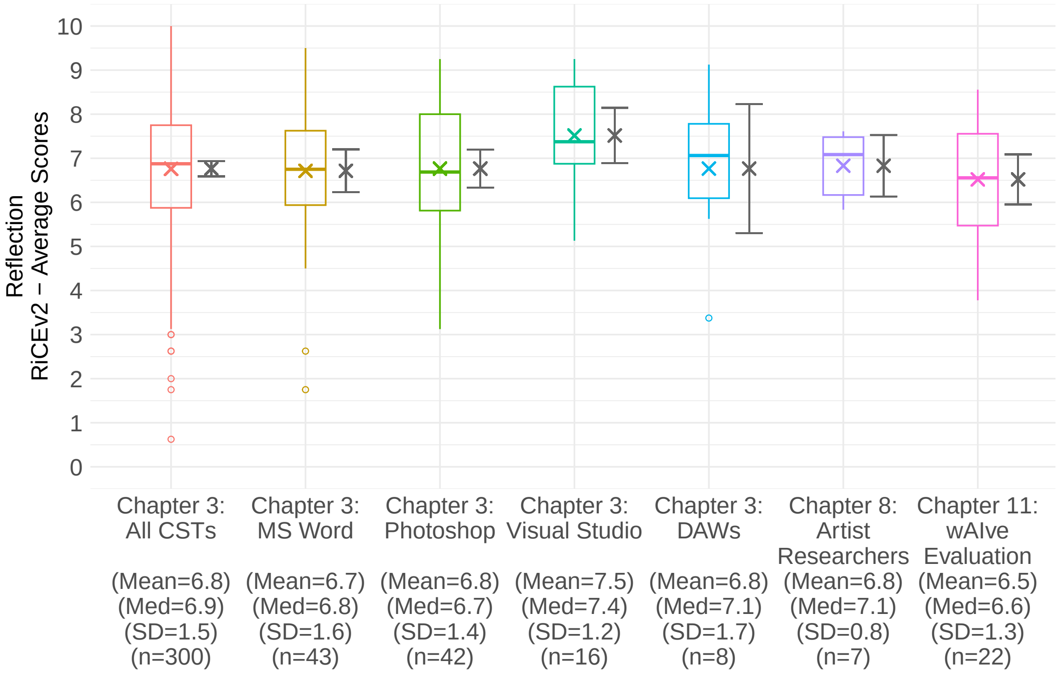


Figure 12.3: RiCEv2 scores for CSTs across chapters.

The highest RiCEv2 scores are from the subset of participants in Chapter 3 who chose Visual Studio. This pattern is also observed for RiCEv2's sub-scales. This shows that programming, or some aspect of programming, prompts more reflection than other domains. Kim and Lerch (1997) show that programming is a cognitively demanding task; thus, it requires more reflection than other creative domains. The iterative nature of writing and executing code also allows for more frequent repetition of reflection cycles. This is similar to the inquiry processes identified in literature on the reflection process (Baumer, 2015; Dewey, 1933).

In the artist-researcher study (Chapter 8), P5, who is a live coder, scored the highest RiCEv2 score (7.8). This supports the notion of programming as a highly reflective activity. Indeed, Sayer (2015) argues that the motor skills of live coding are less challenging to master than for playing an instrument. Thus, the "sensation of being an observer is[...] more vividly conscious" (Sayer, 2015, pg. 3) to live coders than musicians. This explains how reflection during programming, self reported with RiCEv2, is thus more recognisable to live coders.

WAIve has the lowest RiCEv2 score compared with other CSTs. However, the difference in scoring is negligible. The wAIve participants also had the lowest SRIS and MSI scores and were unfamiliar with the interface. This contrasts with the other CSTs in the thesis. WAIve's participants required more time to build their skills (Resnick et al., 2005) before reaching more contemplative states, cf. flow theory (Csíkszentmihályi, 1990). This supports that participants need familiarity with a technology to achieve deeper reflections (Bilda et al., 2008).

With respect to music interaction, comparing wAIve with Chapter 3's subset of DAW users shows that wAIve encouraged less reflection than other music software. As Chapter 11 shows, reflection in wAIve was affected by task disruptions and confusion on how the AI worked. Indeed, Theme 3 from Chapter 11 (*Relationship between focused attention and reflection on AI*) shows that wAIve's flying block animations interrupted people's listening and sometimes caused annoyance. This explains the lower RiCEv2 scoring.

WAIve's users were also not in control of when they were interrupted, contrasting with the artist-researchers in Chapter 8 who chose when to use AI and had few unwanted interruptions. This supports suggestions in HCI to avoid disruption (Adamczyk & Bailey, 2004) and human-centred AI narratives advocating for user control (Shneiderman, 2022). The artist-researchers and DAW users also contrast wAIve in that these users had tacit knowledge (Schön, 1983). This made it easier to reflect on AIGC on instinct, whereas wAIve's novices would not have developed the tacit knowledge required to reflect on AIGC.

Another explanation for wAIve's low scores is that the DAW users self-selected their CST and were thus familiar with music interaction. This contrasts with the music novices using wAIve. There is also a smaller sample for DAWs than wAIve. A stronger comparison is between the DAW users and Chapter 8's artist-researchers: both are musically skilled, and the sample sizes are similar. Many artist-researchers also used DAWs in addition to AIGC in their workflow. The observation that the artist-researcher's scores are similar to the DAW users thus shows that AIGC did not noticeably affect reflection, nor did the differing study settings.

12.3 Reflection-on-Process

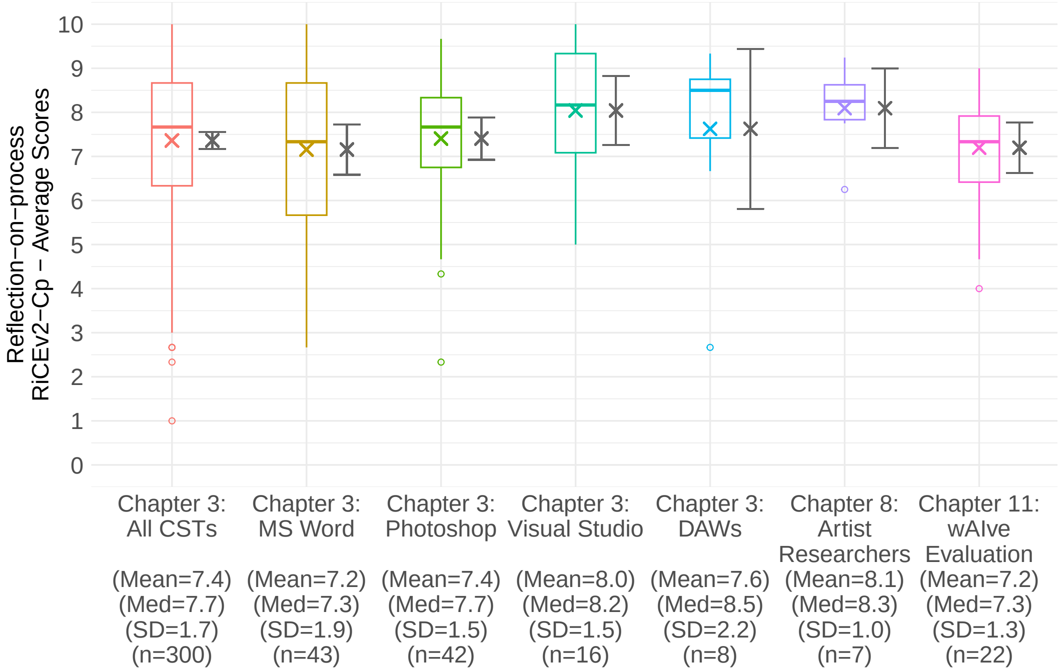


Figure 12.4: RiCEv2-Cp (Reflection-on-process) scores for CSTs across chapters.

Figure 12.4 shows the total RiCEv2-Cp scores for participants' reflection-on-process across CSTs and user studies. A Kruskal-Wallis test (Kruskal & Wallis, 1952) found no significant difference between RiCEv2-Cp scores across the CSTs ($\chi^2 = 6.00$, df = 6, p = .42). However, compared to the

RiCEv2 scores, there is an increase in mean average scores (all between 7.2 and 8.1). Wilcoxon signed-rank tests show significant increases between RiCEv2-Cp and RiCEv2 for: all CSTs in Chapter 3 (V = 33988, $p < .001$), Photoshop (V = 644, p = .03), and the artist-researchers' AI tools (V = 4, $p < .01$). This characterises that participants in each of these cases showed significantly more reflection-on-process than other forms of reflection.

As a wide and diverse sample, the increase for all CSTs in Chapter 3 shows that reflection-on-process contributes more to participants' overall reflection than other types of reflection in CST contexts. This supports the motivation for process-centred CSTs (Sterman, 2022) which encourage either documenting (Dalsgaard & Halskov, 2012; Kim et al., 2017; Sharmin & Bailey, 2013; Sterman et al., 2023) or allowing replay of (Carrera & Lee, 2022; Tseng & Resnick, 2016) the creative process (see Section 2.2.1). Photoshop, compared to the other CSTs, also uses features such as layers and history keeping (Manovich, 2011), allowing users to move back and forth between their edits, supporting reflection on processes over time.

For the artist-researchers, the high reflection-on-process scores are influenced by their use of *reflection boards* (see Section 7.1.4.2), where documenting and reflecting on their composition process was part of the study design. The artist-researchers also reflected-on-process more than in wAIve. They were more familiar with their AI tool and thus better prepared to consider ways to refine their creative process, cf. Bilda et al. (2008).

12.4 Reflection-through-Experimentation

Figure 12.5 shows the total RiCEv2-Ex scores for participants' reflection-through-experimentation across CSTs. A Kruskal-Wallis test (Kruskal & Wallis, 1952) found no significant difference between RiCEv2-Ex scores across the CSTs ($\chi^2 = 2.70$, df = 6, p = .84). A much wider variation in scores is observed than for RiCEv2. There is also no significant difference between the mean RiCEv2 and RiCEv2-Ex scores.

However, the RiCEv2-Ex scores for wAIve are higher than those of artist-researchers. This contrasts with the other RiCEv2 factors, where wAIve scores are lower. The wAIve user study involved participants' assessment of interaction with a novel CST instead of tools they had previously used.

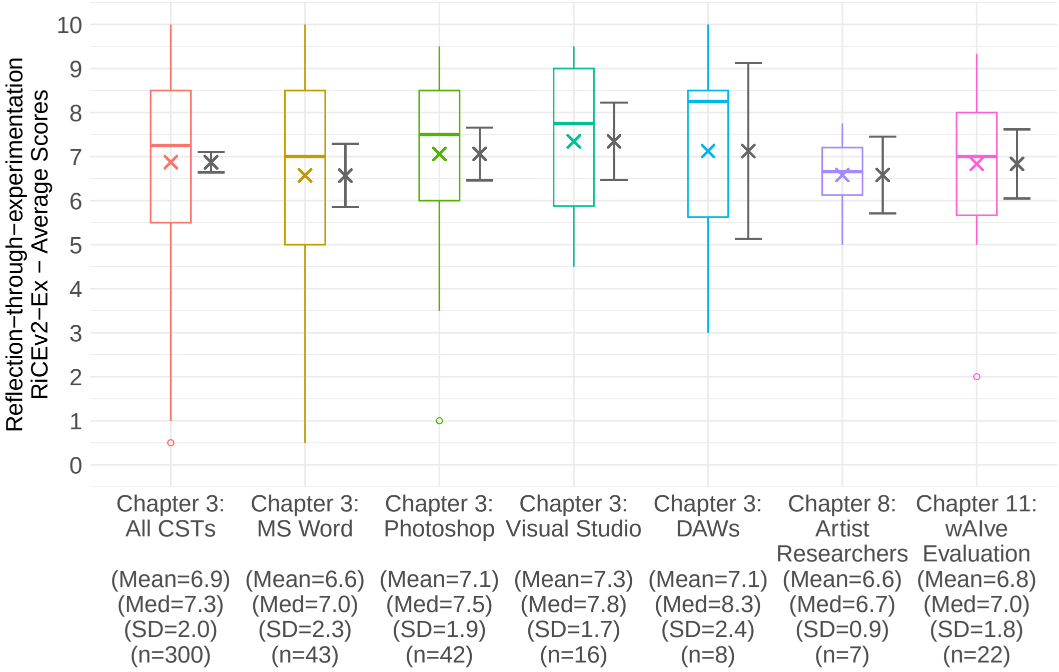


Figure 12.5: RiCEv2-Ex (Reflection-through-experimentation) scores for CSTs across chapters.

There was thus more need to reflect-through-experimentation to learn how wAIve worked. This contrasts with the artist-researchers, whose strong musical skills helped them to experience contemplative states (Csíkszentmihályi, 1990). Instead of contemplative types of reflection, the tinkering required to understand wAIve is represented by the reflection-through-experimentation scores captured by RiCEv2. Exploring wAIve's interface to develop initial skills is also similar to the pattern observed in wAIve's design process (see Chapter 9): early design sessions focused on learning wAIve and improving its usability, before deeper reflection on its design could occur. This corroborates Chapter 8's finding that reflection-through-experimentation is needed early in the composition process (see Figure 8.3).

Furthermore, wAIve included many unfamiliar features to encourage reflection, which targeted reflection-through-experimentation more than other types of reflection. For example, wAIve's animated AI blocks led participants to be reactive to music whilst composing, clicking on blocks as they entered the screen (see Chapter 11 Theme 3). This reacting in-the-making-moment (Candy, 2019) contrasts future-oriented reflections or planning cf.

reflection-on-process (Chapter 3). Similarly, there was evidence of periods of listening and reflection in the artist-researcher's evaluation (see Figure 8.4), which were not identified as prominently in wAIve's evaluation. Where participants in wAIve's evaluation did engage in periods of listening, they tended to be more musically sophisticated and motivated to create quality music (see Chapter 11 Theme 1).

The low reflection-through-experimentation score for the artist-researchers reflects its averaging across composition sessions. The score incorporates later sessions where many participants stopped experimenting and worked on finalising their compositions (see Figure 8.3). WAIve's scores are thus higher because its task length was only 20 minutes, compared to 4 hours. This shows that reflection-through-experimentation is more common in the initial phases of music composition processes (see Figure 8.3). This aligns with models of creative processes, such as the Double Diamond[1]. As a tool, wAIve was also purposefully designed to focus the interaction on music creation processes rather than production processes that require less reflection-through-experimentation (Vanka et al., 2023). For example, see P5 in Figure 8.3 who closed the session by completing standard production processes to finalise their earlier experiments with their AI.

12.5 Reflection-on-Self

Figure 12.6 shows the RiCEv2-Se scores across CSTs for participants' reflection-on-self. A Kruskal-Wallis test (Kruskal & Wallis, 1952) found no significant difference between RiCEv2-Se scores across the CSTs (χ^2 = 9.60, df = 6, p = .14).

Wilcoxon signed-rank tests found significant differences between RiCEv2-Se and RiCEv2 for: all CSTs in Chapter 3 (V = 52860, p < .001), Photoshop (V = 1125, p = .03), and wAIve (V = 327, p < .05). This characterises that reflection-on-self occurred significant less in these cases than other forms of reflection.

As the largest and most diverse sample, the difference for all CSTs from Chapter 3 demonstrates that reflection-on-self is less common in CST interaction in general. This contrasts suggestions from the creative profes-

[1] https://www.designcouncil.org.uk/our-resources/the-double-diamond/

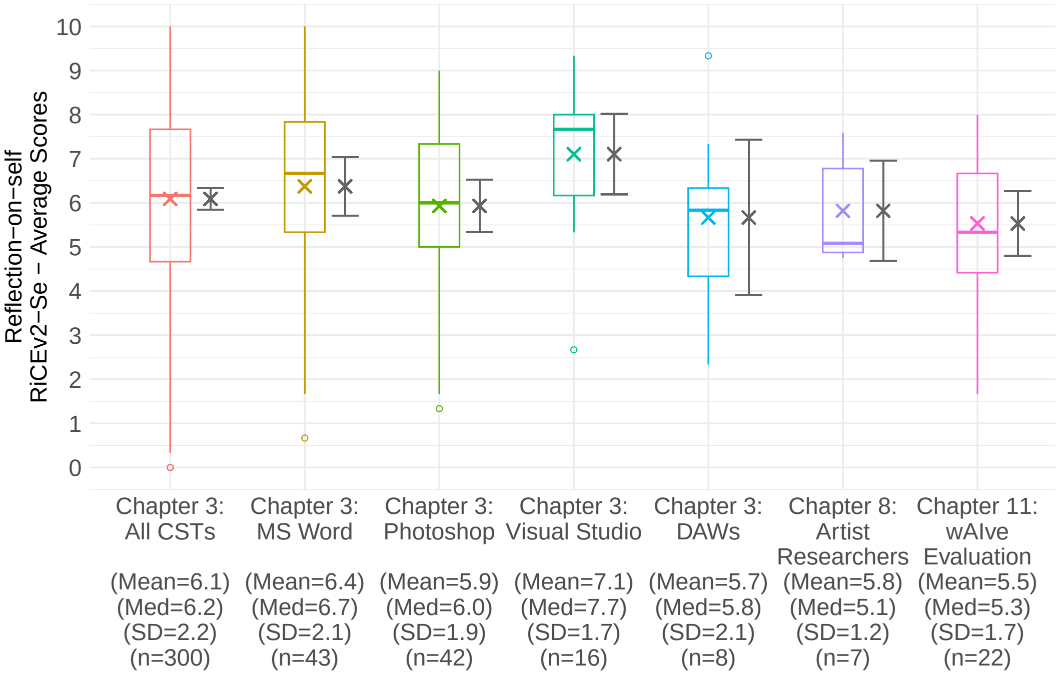


Figure 12.6: RiCEv2-Se (Reflection-on-self) scores for CSTs across chapters.

sionals in Chapter 3 that moments of reflection in creative contexts were linked to "self expression" (P1) and often personal to creators. This also contrasts Sturm et al. (2019), who found AI helped them to find expressions for their self-reflection, and the emotional connections to AI felt by the artist-researchers in Theme 4 from Chapter 7 (*Reflection on feelings*).

WAIve's evaluation participants showed significantly less self-reflection than other types of reflection in their interaction. As above in Section 12.4, this relates to the more reactive interaction style of wAIve, the limited expertise of the participants, and the limited time for the task. The benefits of novice engagement in music composition, such as learning how to question your (musical) intuitions (Kaschub & Smith, 2009) and to develop a musical identity (Barrett & Hickey, 2003; Whittall, 2011), which require self-reflection, are thus not easily accessible to novices. There is no evidence that AIGC thus supported self-reflection. Furthermore, this was evidenced in the case study of the artist-researchers'; self-reflection occurred when organising already curated AIGC and not during people's interaction with the AI model (see Figure 8.4).

The confidence intervals show that the RiCEv2-Se values fall lower for the music-related CSTs than other CSTs. This shows that self-reflection occurs less in musical contexts. This was also demonstrated in the RiCEv1 evaluation (see Chapter 4); RiCEv1-Se scored significantly higher for the task of writing a story with story-sentiment-visualiser, than for the more open-ended interaction with sound-sketcher. Writing is semantic, and thus, it is easier to interpret meanings from (Liu, 2017) than with music. Its semantic nature thus supports self-reflection. Furthermore, Kahneman (2011) characterises reflection-on-self as more cognitively taxing than other types of reflection; reflection-on-self is thus less common as music composition requires more effortful consideration of non-semantic meanings.

12.6 Summary

Table 12.1: Summary of Chapter 12's main findings.

Main Finding	Location
Reflection is characterised as dependent on people's familiarity with a CST tool. Studies where participants were familiar with a CST showed more reflection than other participants.	§12.2 §12.4
WAIve's reflection-through-experimentation scores are higher than other types of reflection, contrasting with other tools. This is informed by participants' low level of familiarity and wAIve's more reactive interaction style.	§12.4
Reflection is characterised in AI-based music composition styles that include programming as highly reflective, with more opportunities for reflection cycles.	§12.2
Reflection-on-process contributes more to people's reflection in CST contexts. Reflection-on-self is less common overall.	§12.3 §12.5
Self-reflection is less common in music interaction than in other creative domains.	§12.5

This chapter inspected the RiCEv2 measures used across CSTs and study settings in this thesis. The main findings are summarised in Table 12.1. The chapter demonstrates that RiCEv2 enables a systematic inspection of reflection in CSTs, showing whether different types of reflection were more or less common for each study tool and interaction context. Based on the findings from this chapter and the study findings throughout, the thesis's research questions are addressed in the following chapter.

Part IV

Conclusions

Chapter 13

Conclusions

This thesis argued that a systematic evaluation of reflection in Creativity Support Tool (CST) interaction is needed to characterise and support the user's creative process. This challenged the dominance of assessments of engagement in the CST field. Indeed, the current state-of-the-art for CSTs was to follow design principles (Resnick et al., 2005; Shneiderman et al., 2006) and evaluation methods (Cherry & Latulipe, 2014) that were based on engagement (Doherty & Doherty, 2018; O'Brien et al., 2018) and flow theory (Csíkszentmihályi, 1990). However, these approaches contrasted with reflection, and there was little consensus on how to operationalise reflection for systematic CST evaluations.

The thesis thus set out to characterise reflection for CST interaction contexts, using the standard for CSTs of mixed-methods evaluations (Hewett et al., 2005). To focus the investigation, Artificial Intelligence (AI)-based music composition was selected as a case study domain in which to characterise reflection; there were few existing user studies on how AI generated music is used (Jourdan & Caramiaux, 2023), and none which directly focused on reflection.

This chapter summarises the thesis's studies and the main findings. It then answers the research questions introduced at the beginning of the thesis (see Section 1.3). What users of AI-based CSTs should thus do is then described in Section 13.1, followed by a review of the thesis' methodological approach and its limitations in Section 13.4. Directions for future work are described to bring the thesis to a close.

Table 13.1: Summary of thesis's main findings.

Main Finding	Location
RQ1: How is reflection characterised in people's open-ended interaction with AI-based CSTs designed for music composition?	
Reflection is characterised in musical AI-based CST interaction using three factors: reflection-on-self, reflection-through-experimentation and reflection-on-process. The factors are measurable through a new self-report questionnaire: the Reflection in Creative Experience (RiCE) questionnaire.	§3.3.5 §5.2.5
Reflection is characterised in CST interaction as including mostly reflection-on-process.	§8.2.1 §8.2.2 §11.2.1 §12.3
Reflection is characterised in CST interaction as including little reflection-on-self, despite self-reflection being judged as important for CST interaction by creative professionals. Musical CST interactions include less self-reflection than in other domains.	§3.2.2.3 §8.2.1 §8.2.2 §11.2.1 §12.5
Reflection is characterised in AI-based music composition as a push and pull between reflection-on-process (when curating AIGC in real time) and reflection-on-self (when arranging already curated AIGC).	§7.2.2.1 Figure 8.4
Reflection is characterised in AI-based music composition as including reflection-through-experimentation early in the creative process. It is often used in unfamiliar situations to develop understandings of AI tools.	§4.2.3 §8.2.1 Figure 8.3 §11.2.3.2
Reflection is characterised in musical AI-based CST interaction as occurring when participants are familiar with an AI tool. This is shown for studies where users are both familiar and unfamiliar with a CST. Users adapt AI to be more familiar to their compositional style, such as by selecting AIGC whilst listening in real-time (as in improvisation).	§7.2.2.1 §12.2 §12.4
Reflection is characterised in AI-based music composition as dependent on users' perceptions of AI, curiosity and motivation. For example, when participants perceive an AI as more skilled than themselves, it is used in lieu of reflection.	§7.2.2.3 §11.2.3.1 §11.2.3.2 §11.2.3.5 §11.2.3.6
RQ2: What is the interplay between characterisations of reflection and engagement in people's open-ended interaction with an AI-based CST for music composition?	
The interplay between reflection and engagement is that moments of reflection occur alongside moments of focused attention.	§11.1.4.2 Figure 11.5
The interplay between reflection and engagement is that self-reflection occurs alongside feelings of a rewarding user experience.	§11.1.4.2 Figure 11.8
The interplay between reflection and engagement is that AIGC is used without reflection in moments of hyperfocus.	§11.2.3.3

13.1 Summary of Thesis and Main Findings

Table 13.1 shows the main findings from the thesis's user studies addressing the research questions introduced at the beginning of this thesis (see Section 1.3). An overview of the thesis studies is given below to remind the reader.

Part I of the thesis addressed that there was no questionnaire for systematically evaluating reflection in CST interaction (see Chapter 2). It developed the Reflection in Creative Experience (RiCE) questionnaire. RiCEv1 was developed by generating a list of 115 items, and reducing these to 4 factors with 8 items through a review of the items by creative professionals and an exploratory factor analysis. Following its evaluation (Chapter 4), RiCEv1 was further updated to RiCEv2 to support its reliability (Chapter 5). RiCEv2 consists of 9 items grouped into three factors of reflection-on-process (reflection on where to take a creative work), reflection-on-self (reflection on personal growth and a creative work's personal meaning) and reflection-through-experimentation (reflection on in-the-moment tinkering and interaction). The factors were conceptually meaningful in that they could successfully show differences in the types of reflection that occurred in different CST interactions and were substantiated by related literature.

With a tool to enable the systematic comparison of reflection in CST interaction developed, Part II of the thesis applied RiCEv2 to the case study domain of AI-based music composition. This addressed that few AI-based CSTs have directly assessed reflection, and none had systematically evaluated reflection across AI-based music composition tools. Chapter 7 thus evaluated reflection in a plurality of music composition styles and AI tools. Seven artist-researchers composed six songs, each using an AI tool of their choice. A mixed-methods approach was used, where participants stopped every hour to reflect on screenshots of their music making and answer the RiCEv2 questions. First-person accounts from the artist-researchers were presented to capture individualistic, subjective and nuanced insights into their composition process. Chapter 8 triangulated the RiCEv2 and qualitative findings to show a characterisation of AIGC in AI-music making (see Figure 8.4). The characterisation is a trade-off relationship between reflection-on-self and reflection-on-process. This addressed research question one by characterising reflection in AI-based music composition as where reflection-on-process occurs whilst curating AIGC by listening in real time, and reflection-on-self occurs whilst organising already curated AIGC.

Part III of the thesis documented the design and development of a novel AI-based CST for music composition, named wAIve. WAIve enabled the evaluation of features common in AI-based music composition and focused on reflection, not additional concerns such as usability (Bryan-Kinns & Reed, 2023). The development and study of wAIve was also justified to reduce variation across findings, in contrast to Part II of the thesis; all participants testing wAIve showed more homogenous characteristics and started with the same level of familiarity. The features developed included animated AI blocks that fly across the user's screen to encourage users to test new music, and flashing play buttons to encourage listening from different perspectives.

Chapter 11 evaluated wAIve by applying RiCEv2 in a mixed-methods user study with 22 computer science students. It was found that higher reflection-on-self scores occurred alongside users with a more rewarding user experience, and that reflection-through-experimentation occurs when people learn something new from observing an AI animation. Animations also broke moments when people were attentive and focused, but depending on the context, were not seen as distracting per se. The findings address the first research question by characterising when reflection on AIGC occurred in the music composition process, such as that reflection-through-experimentation occurred when people learnt new interactions from an AI. The findings also address the second research question of this thesis by characterising the interplay between reflection and engagement for AI-based music composition (see Figure 11.9): for example, that reflection-on-self interplayed with the engagement aspect of reward.

Chapter 12 inspected the RiCEv2 scores across the different CSTs evaluated in the thesis. The main findings were that reflection-on-process was common in CST interaction, whilst reflection-on-self was less common. Reflection-on-self was also less prominent in musical CST domains than in the other CSTs in the thesis. The tools with which participants were familiar were also shown to give higher reflection scores than tools with which participants were less familiar. This answers the first research question by characterising patterns of reflection common in the CSTs, and identifying patterns specific to the AI and music context. For example, self-reflection is characterised as less common in music tools.

13.2 Answers to Research Questions

This section directly answers the research questions introduced at the beginning of this thesis (see Section 1.3).

RQ1: How is reflection characterised in people's open-ended interaction with AI-based CSTs designed for music composition?

Reflection is characterised in people's open-ended interaction with AI-based CSTs designed for music composition as a balance of *reflection-on-process*, *reflection-through-experimentation*, and *reflection-on-self.* These types of reflection were identified by administering RiCEv2 post-hoc to quantify reflection in CST interaction, and triangulating the RiCEv2 scores with qualitative findings and related literature. The types of reflection have temporal qualities and vary throughout the music composition process with AIGC.

Reflection-on-process (reflection on where to take a piece of music) is characterised as the most common type of reflection that occurs in CST interaction. The post-hoc scores of reflection-on-process showed the highest scores compared with the other types of reflection across the thesis' studies on AI-based music interaction. This includes for assessments across AI tools (see Table 8.1) and for the novel CST wAIve (see Figure 11.3).

In interaction with AI-based music CSTs, reflection-on-process is also characterised as commonly occurring when users listen to their music composition or AI generated music. This was shown by the analysis of several artist-researchers use of different AI tools (see Figure 8.4) and qualitative findings from people's interaction with different designs of wAIve (see pg. 171). When listening to the music from different perspectives, the user reflects on how AIGC fits within their compositional aesthetics (see Chapter 7 Theme 1) or how to change their music going forward to fit with the AI tool they are using (such as in Chapter 11 Theme 6 or Chapter 7 Theme 3).

Reflection-on-self is characterised as least common in CST interaction. The assessment of its scores across the CSTs assessed in this thesis shows smaller values in the music interaction domain than for other creative contexts (see Figure 12.6). In AI-based musical CST interaction, reflection-on-self is also characterised as occurring when users were organising already curated AIGC; users would assign meanings to their selected music and consider its personal connection to themselves (see Section 8.3.1). It is notable that reflection-on-self occurs outside of interaction with AIGC; there was no evidence to support that use of AIGC supported or encouraged reflection-on-self (see Section 12.5). Moments of reflection-on-self are further characterised as occurring in a push-and-pull dynamic with reflection-on-process (occurring when curating AIGC), as visualised in Figure 8.4.

Reflection-through-experimentation is characterised as occurring throughout the music composition process. However, it is most prominent at the beginning of composition processes where users are initially learning a tool or setting themselves up to work with the generative capabilities of their selected AI (for example, see Figure 8.3). Furthermore, users often reflect on how an AI works by experimenting. This was shown in wAIve's interaction where users would conduct experiments which mimicked the AI's behaviour (see Chapter 11 Theme 2), or where artist-researchers would tinker with their AI in-the-moment (see Chapter 7 Theme 1).

Reflection-through-experimentation is characterised as more common when users are unfamiliar with an AI. As AI is an emerging technology, reflection-through-experimentation is likely to occur as people use the technology for the first time. Across the user studies in the thesis, participants showed more reflection-through-experimentation with unfamiliar CSTs such as wAIve (see Section 12.4). Furthermore, this is demonstrated by wAIve having many unfamiliar features which led to moments of reflection-through-experimentation, as participants tinkered to learn the new interactions (see Section 12.4).

RQ2: What is the interplay between characterisations of reflection and engagement in people's open-ended interaction with an AI-based CST for music composition?

Interplay is shown between the RiCE characterisations of reflection and the User Engagement Questionnaire (O'Brien et al., 2018) characterisations of engagement in AI-based music composition. This is based on the statistical analysis of wAIve in Chapter 11, which was substantiated through comparisons to related work and triangulation with qualitative findings (see Section 11.3).

There is interplay where moments of reflection occur alongside the engagement aspect of focused attention (see Section 11.3.1.1). In AI-based music composition, however, AI that interrupts users during moments of engagement is both detrimental and conducive to engagement, depending on the context of the interruption and external factors. When users were focused on their music, some found AIGC interruptions broke their creative flow. For example, Section 11.3.2.1 shows that users found AIGC interruptions annoying. Section 11.3.2.1 describes how interruptions led to users performing fast-thinking (Kahneman, 2011) and less reflective modes of interaction to maintain engagement. For example, users would use visual clues in AI generated music, instead of listening to the AIGC, where they would engage in deeper reflection (see Section 11.3.2.1).

There is also interplay between reflection-on-self and the engagement aspect of reward. This is demonstrated in a regression model of the RiCEv2-Se score and UEQ-reward scores (see Figure 11.8). In line with the finding that self-reflection is the least common in interaction with AI-based CST interaction (see pg. 223 above), it is notable that it has interplay with people's feelings of reward. Self-reflection is thus an under-valued aspect of musical CST interaction, as corroborated by the finding that self-reflection is highly valued by creative professionals (see Section 3.2.2).

There is further interplay between reflection and user motivation, which is essential to engagement frameworks such as flow theory (Csíkszentmihályi, 1990). In the case of wAIve, users with a music-first motivation (wanting to create high quality music) were more likely to reflect on AIGC suggestions (see Section 11.3.1.3). Users motivated by a curiosity about how AI worked would also reflect upon AIGC. In contrast, those less interested prefer to perform more fun-based interactions as opposed to engagement (see Section 11.3.2.2). Furthermore, reflection was more likely to occur when users perceived the AI as less skilled than themselves (see Section 11.2.3.5). Those who found AI more skilled than themselves would trust its recommendation without further reflection.

13.3 Recommendations for AI-based Musical CST Users

This section describes how CST users can use the main findings of this thesis for AI-based music composition. In line with the thesis argument (see pg. 219), this section demonstrates how the systematic characterisations of reflection identified in this thesis can be applied to support users in their music composition process. It is recommended that AI-based musical CST users adapt their interaction strategies to intentionally foster different types of reflection as desired.

Recommendation: The AI-based musical CST user should listen to AI generated material in real time to prompt reflection on where to take a piece of music and overcome creative block.

It was found that reflection-on-process occurs when curating AIGC by listening in real-time during composition, whilst reflection-on-self was found to occur when people are organising their already curated AIGC. Thus, if users feel a form of creative block (Lewis, 2023, 2025) and are unsure where to take their music, reflection-on-process can be encouraged by listening to and selecting AI generated music in real time.

Recommendation: The AI-based musical CST user should spend more time organising previously curated AIGC to support self-reflection.

In addition to the finding that reflection-on-self occurs when people are organising their already curated AIGC, the thesis found an interplay between self-reflection and people's feelings of a rewarding experience. Self-reflection was also shown to be less common in musical interaction contexts, despite its benefits. Thus, AI-based CST users should dedicate more time to organising curated content produced by AIGC. This is to support creating music that is rewarding and resonates more with users' personal identities.

Recommendation: The AI-based musical CST user should reflect-through-experimentation to become familiar with an AI tool and its capabilities, to foster further reflections on its outputs.

Reflection-through-experimentation was characterised as common at the start of the creative process, and supports users in developing an understanding of a new AI tool. Therefore, it is recommended that users spend time reflecting-through-experimentation when starting to use an AI-based music composition tool. This will help them develop familiarity with the AI tool and create material that complements its capabilities. Moreover, the finding that the perceived skill of an AI affects reflection (AIs perceived as experts are not reflected upon) supports the notion that reflection-through-experimentation can help users to develop a more nuanced understanding of the AI's capabilities and limitations. Thus, users would be better placed to foster moments of critical reflection when using an AI rather than points of fast thinking (Kahneman, 2011).

Recommendation: The AI-based musical CST user should reflect to learn from AI and its interactions outside moments of focused attention.

The finding that there is interplay between reflection and focused attention, and the conditions under which this occurs, informs how and when users should integrate AIGC into their process. Users should reflect on an AI to understand how it responds and what it produces. This enables them to learn from its interaction patterns and improve their own compositional practice. However, users should refrain from enabling AI agency over their product if they are listening to their music in the moment and in creative flow. This is because hyper-fixated users are likely to ignore or add AIGC without critically engaging with AI generated material.

13.4 Limitations of the Methodological Approach

Table 13.2: Summary of data collection and analysis methods used across the thesis.

Thesis Part	**I**	**II**	**III**
Questionnaires	✓	✓	✓
Descriptive Statistics and Visualisation	✓	✓	✓
Hypothesis Testing	✓	✓	
Regression Analysis		✓	
Interviews	✓	✓	✓
Video-Cued Recall		✓	
Reflection Board			✓
First-person Accounts			✓

This section discusses the methodological approach used in this thesis. It reflexively examines how the data collection and analysis methods contributed to the thesis' findings. Table 13.2 outlines this thesis's data collection and analysis methods. The study task, settings, and interfaces are reflected upon below. This is followed by a discussion of the questionnaires, interviews, and their analysis. Throughout, the limitations are described.

13.4.1 Open-ended Tasks and Settings

This thesis used the open-ended task to "freely compose" a piece of music. This was done to assess qualities characteristic of CST interaction (Kerne et al., 2013). The tasks fit the "how" and "what is" research questions by generating a sufficient breadth of data in which reflection patterns could be characterised. The diversity in data collection also showed findings reflecting a plurality of participants' music composition styles (Biasutti, 2012; McAdams, 2004; McLean & Wiggins, 2010). However, this variation limited the extent to which the answers to the research questions generalise. The findings are entangled with a range of other study factors and should be interpreted within the parameters of the case studies.

Across the studies, the task time varied: 2 minutes to explore the novel interfaces of story-sentiment-visualiser and sound-sketcher in Chapter 3, 15 minutes to explore design iterations of wAIve in Chapter 9, 20 minutes to evaluate wAIve in Chapter 11, and 4 hours to evaluate the AI chosen by participants in Chapters 7 and 8. There were practical reasons for this based on the study settings, the study motivation, and tools used. For example, in the online setting used to develop RiCEv1, participants were paid based on their survey completion time; longer interactions would have been expensive (Müller et al., 2014). It was also helpful to use a long study time to evaluate the range of AI tools in Chapters 7 and 8 because the sample was small and heterogeneous; the longer task afforded multiple rounds of data collection for a richer set of data per participant.

Some of the thesis tools, such as wAIve, provided limited interactions to focus the studies. Longer use of these CSTs would have bored participants (Edmonds, 2014). This boredom would have also confounded findings on engagement and flow states (Csíkszentmihályi, 1990). For wAIve, twenty minutes was found to be effective in providing sufficient insights on engagement, without participants becoming bored. This corroborates timings in other music-related user studies (Bryan-Kinns et al., 2007; Ford et al., 2021). Also, Theme 1 from Chapter 11 (*Participants' motivation informed their reflection*) shows that participants approached tasks more seriously over time. P15 said in "the last 10 minutes[... I] actually tried to do something".

There were other challenges to incorporating engagement into the studies of this thesis beyond wAIve. For example, the artist-researchers were purposefully interrupted throughout their music composition, posing a confounding variable for engagement. Assessing engagement would been inappropriate, inflated the project's scope, and added time to the already lengthy study procedure.

13.4.2 Study Tools

The studies investigated various CSTs, with different levels of familiarity for participants. For example, participants self-selected tools they had used before in Chapter 3, and Chapters 7 and 8. All participants were unfamiliar with wAIve. This familiarity influences findings, as people more familiar with an interface show more creative thinking (Bilda et al., 2008). The range of CSTs in this thesis also limited the homogeneity of the data collected throughout the thesis. A more selective approach would have focused the study on tools with similar aspects (such as tools that modify timbre or only programming-based tools). However, this would have also limited the breadth of the findings. It thus would not have been possible to identify which of the tools assessed in this thesis have the most potential to encourage reflection. For example, it was found that live coding interfaces show strength in encouraging reflection.

13.4.3 Quantitative: Questionnaires

The self-report questionnaires allowed for the systematic inspection of different participants and CSTs. Indeed, the Goldsmiths MSI (Müllensiefen et al., 2014) helped to show variation in musical expertise, which was not obvious from the recruitment process. For example, the wAIve evaluation study attracted participants with good musical expertise even though this was not specified in the advert for participation. The MSI accounted for this confounding factor. The SRIS scores (Grant et al., 2002) were useful in describing if there was a confounding influence impacting the RICEv2 scores (Bentvelzen et al., 2022), where participants with a natural reflective capacity would have more generously evaluated the CSTs in this thesis.

The CSI (Cherry & Latulipe, 2014) could have been used throughout studies to connect more closely with other CST research. However, the more targeted UEQ (O'Brien et al., 2018) measure was shorter to complete. This left study time to capture more individualistic data from lengthier qualitative procedures such as video-cued recall (Candy et al., 2006; Candy, 2006). The unrelated factors from the CSI would also not have been beneficial in answering the thesis research questions.

13.4.4 Quantitative: Questionnaire Analysis

Development of the RICEv2 questionnaire was central to the contribution of this thesis. However, there were several challenges to adopting a questionnaire design approach to evaluate creative interactions. For the development of questionnaires, HCI research follows a standard procedure (Boateng et al., 2018), where several binary choices of statistical criteria must be met to decide if a questionnaire is valid or not. This favours data with restricted statistical properties. However, homogeneity is not characteristic of open-ended creative contexts. For example, one aspect of confirmatory factor analysis (see Section 4.1.4.1) is to assess whether different factors in a questionnaire are distinct and do not overlap. Striving to achieve no overlap between types of reflection is challenging as it is an abstract and multifaceted concept (Baumer, 2015; Baumer et al., 2014; Bentvelzen et al., 2022; Fleck & Fitzpatrick, 2010). Homogeneous data tends to lead to more favourable statistics and is achieved through large sample sizes, more precise selection of participants, or a more focused study task (Dix, 2020). These qualities detract from ecological validity when examining CST interaction, which is more nuanced and varied.

The regression model analysis in Section 11.2.2 was limited by noise in the user study data. While this was controlled by analysing confounding variables (see Table 11.3), this limits the generalisability of the models. Given this limitation, the descriptive statistics and visualisations were most helpful in analysing the questionnaires. For example, the visualisations in Figures 8.1 through 8.3 showed findings on how reflection varied across different creative processes. In particular, using different shapes to highlight individual perspectives helped to observe patterns for different groupings of participants based on their qualitative data. The individual points could be mapped to the interviews and artist-researchers' discussions of their composition as it unfolded over time, supporting triangulation of the data. However, the visualisation approach to analysis has limited repeatability; observations are based on the researcher's intuition and not statistically determined metrics such as p-values.

13.4.5 Qualitative: Interviews and First-Person Accounts

The interview approaches used across this thesis helped to understand how people reflected on their interactions. The questions used varied in openness, depending on the study context and goals. For example, the questions used to discuss the iterative design of wAIve were mapped to design goals. At points, this led participants towards certain responses. For example, by asking "Did any parts of the interface encourage you to reflect?", participants could have rationalised a moment of reflection with the interface post-hoc. However, this was justified for the design of wAIve, as participants were briefed on the design goals upfront and explicitly asked to consider these in their interaction.

Video-cued recall (Candy et al., 2006; Candy, 2006) was used to evaluate wAIve in Chapter 11, where the interview questions shifted focus to probing participants' thinking during their interaction. Participants were invited to stop the video and comment at any time. However, the researcher had to leverage the semi-structured approach to collect sufficient detail. The participants were novices in music composition, lacking experience discussing musical interaction. Therefore, the participants needed prompting to elaborate on the study goals. Nonetheless, video-replay gave a focal point for discussion; participants who were less expressive could use the video to guide their discussion.

In Chapter 7, the interviews were centred around reflection boards (see Section 7.1.4.2). In contrast to the video-cued recall approach, this allowed participants to reflect on their composing privately and to consider their interaction more deeply. The interview questions were thus more open-ended, asking participants to talk through their reflections. The approach was also more objective, as the interviewer imposed minimal influence on the process. However, as the interviewer had less impact on the direction of discussion, participants followed tangents distracting from the study's focus on reflection. This led to some unsurprising findings, such as on aspects of the technical setup (see Chapter 7 Theme 6). The reflection boards also required stopping to think throughout the composition process, creating an interruption which confounded studying aspects of engagement such as flow states (Csíkszentmihályi, 1990).

13.4.6 Qualitative: Thematic Analysis

The reflexive thematic analysis (Braun & Clarke, 2006; Braun & Clarke, 2019) provided a repeatable set of steps for analysing qualitative data. Specifically, the *inductive* approach systematised the analysis whilst allowing flexibility. For example, the researcher could account for moments in the semi-structured interview approach where discussion followed tangents. The inductive approach also allowed the researcher to continually revisit the data until themes relevant to the research questions were developed in sufficient depth. However, balancing depth of insight with relevance to the research questions was challenging, as the interpretation of the depth of a theme is essentially a value judgement. A different approach would have been to adapt the factors from the RiCE questionnaire as a deductive coding scheme (or adopt existing coding schemes such as Hubbard et al. (2023)), to support more direct triangulation with the quantitative measures.

The findings generated from this thesis's reflexive thematic analysis approach were limited. Firstly, by dividing transcripts into codes and later rearranging these codes, the temporal aspects of the data were removed. Indeed, the video-cued recall approach and first-person interviews led to data describing linearly how the process of a music composition unfolds. However, these were lost through the coding procedure. Second, many of the findings of the thematic analysis were predictable, such as Chapter 11 Theme 1 (*Participants' motivation informed their reflection*) and Chapter 7 Theme 6 (*Reflection on Technical Challenges*). Whilst leading to more nuanced insights than questionnaire measures, which reduce concepts into categories, the novelty of the qualitative findings arose from their triangulation with the questionnaire measures, such as in Figures 11.10 and Figure 8.4.

The first-person produced the richest detail and was most personal (Ellis et al., 2011; Fdili Alaoui, 2023). They also added to the levels of detail in the analysis; there was detail for each participant (first-person accounts) and the sample (thematic analysis). This gave flexibility in identifying interesting qualities about individuals whilst commenting on commonalities within the sample, which is not captured by thematic analysis alone. These different levels of detail could have been extracted by qualitative analysis methods that highlight differences between participants, such as Diffractive Analysis (Morrison & McPherson, 2024; Nordmoen & McPherson, 2022; Rajcic et al., 2024; Robson et al., 2024). However, such methods are yet to be standardised and are not repeatable. Overall, the first-person accounts helped interrogate reflection as participants were given space to articulate their process (Edmonds, 2022).

13.5 Future Work

Other CST researchers can use RiCEv2 to evaluate if its types of reflection are present in different tools. However, outside of AI-based music composition, the reliability of RICEv2 has not been tested. Studies must be conducted for different creative domains to assess RiCEv2's reliability. The state-of-the-art test is for a confirmatory factor analysis to confirm the structural properties of RiCEv2 and establish that its factors are conceptually distinct and do not overlap (Kline, 2015). As discussed in Section 13.4.4, this is challenging in creative research where reflection is abstract and not easily separated into distinct categories. Future work could crowdsource datasets from researchers who adopt RiCEv2 for their user studies and to show benchmark scores for each domain. There is also an opportunity to translate the RiCEv2 questionnaire to be studied outside of English-speaking countries.

The findings showed that live coding presents high levels of reflection (see Figure 12.3). There are interesting parallels between modes of intuitive thinking whilst programming in real-time and more deliberate modes of thinking, which could extend this thesis's investigation on reflection and engagement (Sayer, 2015). There is also a diverse range of live coding languages (Aaron & Blackwell, 2013; McLean & Wiggins, 2010; Wilson et al., 2021) with their own language primitives. Future work can investigate how these languages' nuances influence reflection. The challenge would be studying live coding in an ecologically valid setting, and evaluating different ways of capturing aspects of reflection and engagement during performance.

This thesis focused on operationalising reflection for CST user studies. Several other aspects of human-AI interactions have yet to be operationalised for the systematic evaluation of CSTs. For example, agency is often a key concern when using AIGC in arts contexts (Amershi et al., 2019; Boden & Edmonds, 2009; Lewis, 2023; Louie et al., 2020; Wilson et al., 2023; Xambó, 2022). Furthermore, aesthetic qualities that artists capture in their AI-inspired artworks – ranging from cute (Medley et al., 2020) to creepy (Woźniak et al., 2021) – also have yet to be examined systematically. Future research can develop methodological tools for these experiential qualities. There is an opportunity to combine these approaches into an evaluation toolkit for human-AI interaction in creative contexts and move towards a more arts and human-centred approach to AI evaluation. This echoes calls to embrace less common user experience evaluation strategies in NIME (Reimer & Wanderley, 2021) and across CSTs (Cox et al., 2025).

This thesis is situated within the field of HCI and the CST subfield. Yet, the research touches on the related areas of Computational Creativity (Colton & Wiggins, 2012) and NIME (Poupyrev et al., 2001). These fields could examine reflection from alternative perspectives. For example, computational creativity research has considered ways that computers could be reflective by analysing and re-writing their code and algorithms (Pérez & Sharples, 2001; Wiggins, 2006). Researchers could thus investigate whether computational approaches lead to novel insights into the nature of reflection, or whether computationally creative AI tools encourage more or less reflection in people's music making. There is also opportunity to explore reflection beyond the case study area of this thesis, for example, in dance (Fdili Alaoui, 2019) or sketching (Lewis et al., 2023). The methodological approach in Chapters 7 and 8 in particular would be interesting to apply to a wide range of domains, to showcase more pluralistic approaches to how people reflect in different creative user experiences – expanding the investigation beyond traditional HCI methods.

References

Aaron, S., & Blackwell, A. F. (2013). From Sonic Pi to Overtone: Creative Musical Experiences with Domain-specific and Functional Languages. *Proceedings of the First ACM SIGPLAN Workshop on Functional Art, Music, Modeling & Design*, 35–46. https://doi.org/10.1145/2505341.2505346

Adamczyk, P. D., & Bailey, B. P. (2004). If Not Now, When? The Effects of Interruption at Different Moments within Task Execution. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 271–278. https://doi.org/10.1145/985692.985727

Addessi, A. R., Ferrari, L., & Carugati, F. (2015). The Flow Grid: A Technique for Observing and Measuring Emotional State in Children Interacting with a Flow Machine. *Journal of New Music Research*, *44*(2), 129–144. https://doi.org/10.1080/09298215.2014.991738

Albert, R. S., & Runco, M. A. (1999). A History of Research on Creativity. In R. J. Sternberg (Ed.), *Handbook of Creativity* (pp. 16–34). Cambridge University Press.

Amabile, T. M. (1982). Social Psychology of Creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology*, *43*(5), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997

Amabile, T. M. (1983). The Social Psychology of Creativity: A Componential Conceptualization. *Journal of Personality and Social Psychology*, *45*(2), 357–376. https://doi.org/10.1037/0022-3514.45.2.357

Amabile, T. M. (1996). *Creativity In Context: Update To The Social Psychology Of Creativity*. Westview Press. https://doi.org/10.4324/9780429501234

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

An, P., Bakker, S., Ordanovski, S., Taconis, R., Paffen, C. L., & Eggen, B. (2019). Unobtrusively Enhancing Reflection-in-Action of Teachers through Spatially Distributed

Ambient Information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300321

Atkins, S., & Murphy, K. (1993). Reflection: A Review of the Literature. *Journal of Advanced Nursing*, *18*(8), 1188–1192. https://doi.org/10.1046/j.1365-2648.1993.18081188.x

Banar, B., & Colton, S. (2021). Generating Music with Extreme Passages using GPT-2. In A. Mora & E.-A. A.I. (Eds.), *Evo* 2021 Late Breaking Abstracts* (pp. 31–35). https://arxiv.org/pdf/2106.11804.pdf

Banar, B., & Colton, S. (2022). A Systematic Evaluation of GPT-2-Based Music Generation. In T. Martins, N. Rodríguez-Fernández, & S. M. Rebelo (Eds.), *Artificial Intelligence in Music, Sound, Art and Design* (pp. 19–35). Springer International Publishing.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Barrett, M. S., & Hickey, M. (2003). Freedoms and Constraints: Constructing Musical Worlds through the Dialogue of Composition. In *Composition in the Schools: A New Horizon for Music Education* (pp. 3–27). MENC.

Bartlett, M. S. (1950). Tests of Significance in Factor Analysis. *British Journal of Statistical Psychology*, *3*(2), 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Baumer, E. P. (2015). Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 585–594. https://doi.org/10.1145/2702123.2702234

Baumer, E. P., Khovanskaya, V., Matthews, M., Reynolds, L., Schwanda Sosik, V., & Gay, G. (2014). Reviewing Reflection: On the Use of Reflection in Interactive System Design. *Proceedings of the 2014 Conference on Designing Interactive Systems*, 93–102. https://doi.org/10.1145/2598510.2598598

Belakova, J., & Mackay, W. E. (2021). SonAmi: A Tangible Creativity Support Tool for Productive Procrastination. *Proceedings of the 13th Conference on Creativity and Cognition*. https://doi.org/10.1145/3450741.3465250

Bellingham, M. (2022, February). *Choosers: A Visual Programming Language for Nondeterministic Music Composition by Non-Programmers* [Doctoral dissertation, The Open University]. https://oro.open.ac.uk/81935/

Benford, S., Greenhalgh, C., Crabtree, A., Flintham, M., Walker, B., Marshall, J., Koleva, B., Rennick Egglestone, S., Giannachi, G., Adams, M., Tandavanitj, N., & Row Farr, J. (2013). Performance-Led Research in the Wild. *ACM Transations in Computer-Human Interaction*, *20*(3). https://doi.org/10.1145/2491500.2491502

Benjamin, J. J., Biggs, H., Berger, A., Rukanskaitė, J., Heidt, M. B., Merrill, N., Pierce, J., & Lindley, J. (2023). The Entoptic Field Camera as Metaphor-Driven Research-through-Design with AI Technologies. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581175

Bennett, S. (1976). The Process of Musical Creation: Interviews with Eight Composers. *Journal of Research in Music Education*, *24*(1), 3–13. https://doi.org/10.2307/3345061

Ben-Tal, O., Harris, M. T., & Sturm, B. L. (2021). How Music AI Is Useful: Engagements with Composers, Performers and Audiences. *Leonardo*, *54*(5), 510–516. https://doi.org/10.1162/leon_a_01959

Bentler, P. M., & Bonett, D. G. (1980). Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Bentvelzen, M., Niess, J., Woźniak, M. P., & Woźniak, P. W. (2021). The Development and Validation of the Technology-Supported Reflection Inventory. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445673

Bentvelzen, M., Woźniak, P. W., Herbes, P. S., Stefanidi, E., & Niess, J. (2022). Revisiting Reflection in HCI: Four Design Resources for Technologies That Support Reflection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, *6*(1). https://doi.org/10.1145/3517233

Biasutti, M. (2012). Group Music Composing Strategies: A Case Study Within a Rock Band. *British Journal of Music Education*, *29*(3), 343–357. https://doi.org/10.1017/S0265051712000289

Bilda, Z., Edmonds, E., & Candy, L. (2008). Designing for Creative Engagement. *Design Studies*, *29*(6), 525–540. https://doi.org/10.1016/j.destud.2008.07.009

Blackwell, A. F., & Green, T. R. G. (2000). A Cognitive Dimensions Questionnaire Optimised for Users. *Proceedings of the Workshop on Philosophy of Programming Interest Group (PPIG)*, 137–154. https://ppig.org/files/2000-PPIG-12th-blackwell.pdf

Blum, S. (2001). Composition. In *Grove Music Online*. Oxford University Press. https://doi.org/10.1093/gmo/9781561592630.article.06216

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Boden, M. A. (1991). *The Creative Mind: Myths and Mechanisms*. Basic Books, Inc.

Boden, M. A., & Edmonds, E. A. (2009). What is Generative Art? *Digital Creativity*, *20*(1-2), 21–46. https://doi.org/10.1080/14626260902867915

Bødker, S. (2015). Third-Wave HCI, 10 Years Later—Participation and Sharing. *Interactions*, *22*(5), 24–31. https://doi.org/10.1145/2804405

Bono, E. D. (1985). *Six Thinking Hats*. Little, Brown & Company.

Botella, M., Nelson, J., & Zenasni, F. (2019). It Is Time to Observe the Creative Process: How to Use a Creative Process Report Diary (CRD). *Journal of Creative Behavior*, *53*(2), 211–221. https://doi.org/10.1002/jocb.172

Boud, D., Keogh, R., & Walker, D. (1985). Promoting Reflection in Learning: A Model. In D. Boud, R. Keogh, & D. Walker (Eds.), *Reflection: Turning Experience into Learning* (pp. 19–40). Nichols Publishing Company.

Bowman, N. D., Lin, J. T., & Wu, C. (2021). A Chinese-Language Validation of the Video Game Demand Scale (VGDS-C): Measuring the Cognitive, Emotional, Physical, and Social Demands of Video Games. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445348

Boyd, E. M., & Fales, A. W. (1983). Reflective Learning: Key to Learning from Experience. *Journal of Humanistic Psychology*, *23*(2), 99–117. https://doi.org/10.1177/002216788323201

Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2013). *Successful Qualitative Research: A Practical Guide for Beginners*. SAGE Publications.

Braun, V., & Clarke, V. (2019). Reflecting on Reflexive Thematic Analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Braun, V., & Clarke, V. (2021). One Size Fits All? What Counts as Quality Practice in (Reflexive) Thematic Analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Briot, J.-P., Hadjeres, G., & Pachet, F.-D. (2017). *Deep Learning Techniques for Music Generation*. Springer Cham. https://doi.org/10.1007/978-3-319-70163-9

Bryan-Kinns, N., Banar, B., Ford, C., Reed, C. N., Zhang, Y., Colton, S., & Armitage, J. (2021). Exploring XAI for the Arts: Explaining Latent Space in Generative Music. *1st Workshop on eXplainable AI Approaches for Debugging and Diagnosis*, 14. https://xai4debugging.github.io/files/papers/exploring_xai_for_the_arts_exp.pdf

Bryan-Kinns, N., Fiebrink, R., Perry, P., Wilson, E., & Wszeborowska, A. (2024). Responsible AI Music Artistic Mini-Projects. https://ualresearchonline.arts.ac.uk/id/eprint/23595/

Bryan-Kinns, N., Ford, C., Chamberlain, A., Benford, S. D., Kennedy, H., Li, Z., Qiong, W., Xia, G. G., & Rezwana, J. (2023). Explainable AI for the Arts: XAIxArts. *Proceedings of the 15th Conference on Creativity and Cognition*, 1–7. https://doi.org/10.1145/3591196.3593517

Bryan-Kinns, N., & Hamilton, F. (2012). Identifying Mutual Engagement. *Behaviour & Information Technology*, *31*(2), 101–125. https://doi.org/10.1080/01449290903377103

Bryan-Kinns, N., Healey, P. G. T., & Leach, J. (2007). Exploring Mutual Engagement in Creative Collaborations. *Proceedings of the 6th ACM SIGCHI Conference on Creativity & Cognition*, 223–232. https://doi.org/10.1145/1254960.1254991

Bryan-Kinns, N., Noel-Hirst, A., & Ford, C. (2024). Using Incongruous Genres to Explore Music Making with AI Generated Content. *Proceedings of the ACM Conference on Creativity and Cognition (C&C)*. https://doi.org/10.1145/3635636.3656198

Bryan-Kinns, N., & Reed, C. N. (2023). A Guide to Evaluating the Experience of Media and Arts Technology. In A. L. Brooks (Ed.), *Creating Digitally: Shifting Boundaries: Arts and Technologies—Contemporary Applications and Concepts* (pp. 267–300). Springer International Publishing. https://doi.org/10.1007/978-3-031-31360-8_10

Bryan-Kinns, N., Wang, W., & Wu, Y. (2018). Thematic Analysis for Sonic Interaction Design. *Proceedings of British HCI 2018*, 1–3. https://doi.org/10.14236/ewic/hci2018.214

Burnard, P. (1995). Task Design and Experience in Composition. *Research Studies in Music Education*, *5*(1), 32–46. https://doi.org/10.1177/1321103X9500500104

Burnard, P. (2000). Examining Experiential Differences between Improvisation and Composition in Children's Music-making. *British Journal of Music Education*, *17*(3), 227–245. https://doi.org/10.1017/S0265051700000310

Cai, X., Cebollada, J., & Cortiñas, M. (2022). Self-Report Measure of Dispositional Flow Experience in the Video Game Context: Conceptualisation and Scale Development. *International Journal of Human-Computer Studies*, *159*. https://doi.org/https://doi.org/10.1016/j.ijhcs.2021.102746

Caillon, A., & Esling, P. (2021). RAVE: A Variational Autoencoder for Fast and High-quality Neural Audio Synthesis. *arXiv preprint*. https://arxiv.org/abs/2111.05011

Caine, K. (2016). Local Standards for Sample Size at CHI. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 981–992. https://doi.org/10.1145/2858036.2858498

Candy, L., Amitani, S., & Bilda, Z. (2006). Practice-led Strategies for Interactive Art Research. *CoDesign*, *2*(4), 209–223. https://doi.org/10.1080/15710880601007994

Candy, L. (2006). Practice Based Research: A Guide. *Creativity & Cognition Studios (CCS) Report*, *1*(2), 1–19.

Candy, L. (2011). Research and Creative Practice. In E. A. Edmonds & L. Candy (Eds.), *Interacting: Art, Research and the Creative Practitioner* (1st, pp. 33–59). Libri Publishing UK.

Candy, L. (2019). *The Creative Reflective Practitioner: Research Through Making and Practice* (1st). Routledge. https://doi.org/10.4324/9781315208060

Candy, L., & Edmonds, E. (1999). Introducing Creativity to Cognition. *Proceedings of the Third Conference on Creativity and Cognition*, 3–6. https://doi.org/10.1145/317561.317562

Candy, L., & Edmonds, E. (2018). Practice-based Research in the Creative Arts: Foundations and Futures from the Front Line. *Leonardo*, *51*(1), 63–69. https://doi.org/10.1162/LEON_a_01471

Caramiaux, B., & Fdili Alaoui, S. (2022). "Explorers of Unknown Planets": Practices and Politics of Artificial Intelligence in Visual Arts. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2). https://doi.org/10.1145/3555578

Caramiaux, B., Lotte, F., Geurts, J., Amato, G., Behrmann, M., Bimbot, F., Falchi, F., Garcia, A., Gibert, J., Gravier, G., Holken, H., Koenitz, H., Lefebvre, S., Liutkus, A., Perkis, A., Redondo, R., Turrin, E., Viéville, T., & Vincent, E. (2019, April). *AI in the Media and Creative Industries* (Research Report). New European Media (NEM). https://inria.hal.science/hal-02125504

Carnovalini, F., & Rodà, A. (2020). Computational Creativity and Music Generation Systems: An Introduction to the State of the Art. *Frontiers in Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00014

Carrera, D., & Lee, S. W. (2022). Watch Me Write: Exploring the Effects of Revealing Creative Writing Process through Writing Replay. *Proceedings of the 14th Conference on Creativity and Cognition*, 146–160. https://doi.org/10.1145/3527927.3532806

Chamberlain, A., Crabtree, A., Rodden, T., Jones, M., & Rogers, Y. (2012). Research in the Wild: Understanding 'in the wild' Approaches to Design and Development. *Proceedings of the Designing Interactive Systems Conference*, 795–796. https://doi.org/10.1145/2317956.2318078

Chang, M., Druga, S., Fiannaca, A. J., Vergani, P., Kulkarni, C., Cai, C. J., & Terry, M. (2023). The Prompt Artists. *Proceedings of the 15th Conference on Creativity and Cognition*, 75–87. https://doi.org/10.1145/3591196.3593515

Chapman, P. M. (1997). *Models of Engagement: Intrinsically Motivated Interaction with Multimedia Learning Software* [PhD Thesis]. University of Waterloo.

Charnley, J., Pease, A., & Colton, S. (2012). On the Notion of Framing in Computational Creativity. *Proceedings of the 3rd International Conference on Computational Creativity*. https://computationalcreativity.net/iccc2012/wp-content/uploads/2012/05/077-Charnley.pdf

Chen, J. C. W., & O'Neill, S. A. (2020). Computer-mediated Composition Pedagogy: Students' Engagement and Learning in Popular Music and Classical Music. *Music Education Research*, *22*(2), 185–200. Retrieved August 16, 2023, from https://doi.org/10.1080/14613808.2020.1737924

Cherry, E., & Latulipe, C. (2014). Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Transactions on Computer-Human Interaction*, *21*(4). https://doi.org/10.1145/2617588

Cho, Y.-T., Kuo, Y.-L., Yeh, Y.-T., Liang, H.-H., & Li, Y.-T. (2022). Motion-centric Tools to Reflect on Digital Creative Experiences and Created Outputs. *Proceedings of the 14th Conference on Creativity and Cognition*, 234–246. https://doi.org/10.1145/3527927.3531454

Chowning, J. (1977). Stria [Commissioned by IRCAM (Paris) for the Institute's first major concert series: Perspectives of the 20th Century].

Collins, D. (2007). Real-time Tracking of the Creative Music Composition Process. *Digital Creativity*, *18*(4), 239–256. https://doi.org/10.1080/14626260701743234

Colton, S., Pease, A., Guckelsberger, C., McCormack, J., & Llano, M. T. (2020). On the Machine Condition and its Creative Expression. *Proceedings of the International Conference on Computational Creativity (ICCC)*, 342–349. http://computationalcreativity.net/iccc20/papers/ICCC20%7B%5C_%7DProceedings.pdf

Colton, S., Pease, A., & Saunders, R. (2018). Issues of Authenticity in Autonomously Creative Systems. *Proceedings of the 9th International Conference on Computational Creativity*. https://computationalcreativity.net/iccc2014/wp-content/uploads/2014/06/15.8_Colton.pdf

Colton, S., & Ventura, D. (2014). You Can't Know my Mind: A Festival of Computational Creativity. *Proceedings of the 5th International Conference on Computational Creativity*.

Colton, S., & Wiggins, G. A. (2012). Computational Creativity: The Final Frontier? *Proceedings of the 20th European Conference on Artificial Intelligence*, 21–26. https://computationalcreativity.net/iccc2014/wp-content/uploads/2013/09/ComputationalCreativity.pdf

Compton, K., & Mateas, M. (2015). Casual Creators [event-place: Park City, Utah, USA]. *Proceedings of the 6th International Conference on Computational Creativity*, 228–235. https://computationalcreativity.net/iccc2015/proceedings/10_2Compton.pdf

Compton, K. (2019). *Casual Creators: Defining A Genre of Autotelic Creativity Support Systems* [PhD Thesis]. University of California Santa Cruz. https://escholarship.org/uc/item/4kg8g9gd

Cornock, S., & Edmonds, E. (1973). The Creative Process Where the Artist Is Amplified or Superseded by the Computer. *Leonardo*, *6*(1), 11–16. http://www.jstor.org/stable/1572419

Costello, B., & Edmonds, E. (2007). A Study in Play, Pleasure and Interaction Design. *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces*, 76–91. https://doi.org/10.1145/1314161.1314168

Cox, S. R., Djernæs, H. B., & van Berkel, N. (2025). Beyond Productivity: Rethinking the Impact of Creativity Support Tools [event-place: New York, NY, USA]. *Creativity and Cognition (C&C '25)*, 15. https://doi.org/10.1145/3698061.3726924

Crabtree, A., Rouncefield, M., & Tolmie, P. (2012). Design Ethnography in a Nutshell. In A. Crabtree, M. Rouncefield, & P. Tolmie (Eds.), *Doing Design Ethnography* (pp. 183–205). Springer London. https://doi.org/10.1007/978-1-4471-2726-0_10

Creswell, J. W. (2009). Mixed Methods Procedures. In *Research design: Qualitative, Quantitative, and Mixed Methods Approaches* (3rd, pp. 188–206). Sage Publications, Inc.

Croft, J. (2007). Theses On Liveness. *Organised Sound*, *12*(1), 59–66. https://doi.org/10.1017/S1355771807001604

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cross, N. (2008). *Engineering Design Methods.* Wiley.

Csíkszentmihályi, M. (1990). *Flow: The Psychology of Optimal Experience.* Harper Collins.

Csíkszentmihályi, M. (1999). Implications of a Systems Perspective for the Study of Creativity. In R. J. Sternberg (Ed.), *Handbook of Creativity* (pp. 313–338). Cambridge University Press.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv preprint.* https://doi.org/https://doi.org/10.48550/arXiv.1901.02860Focustolearnmore

Dalsgaard, P., & Halskov, K. (2012). Reflective Design Documentation. *Proceedings of the Designing Interactive Systems Conference*, 428–437. https://doi.org/10.1145/2317956.2318020

Dannemann, T., & Barthet, M. (2021). SonicDraw: A Web-based Tool for Sketching Sounds and Drawings. *Proceedings of the International Computer Music Conference 2021*, 301–308. https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/88517/Dannemann%20SonicDraw:%20a%20web-based%20tool%

20for%20sketching%20sounds%20and%20drawings%202021%20Published.pdf?sequence=2

Dannemann, T., Bryan-Kinns, N., & McPherson, A. (2023, May). Self-Sabotage Workshop: A Starting Point to Unravel Sabotaging of Instruments as a Design Practice. In M. Ortiz & A. Marquez-Borbon (Eds.), *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 70–78). http://nime.org/proceedings/2023/nime2023_9.pdf

Davis, N. (2021). Human-Computer Co-Creativity: Blending Human and Computational Creativity. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *9*(6), 9–12. https://doi.org/10.1609/aiide.v9i6.12603

Deterding, S., Hook, J., Fiebrink, R., Gillies, M., Gow, J., Akten, M., Smith, G., Liapis, A., & Compton, K. (2017). Mixed-Initiative Creative Interfaces. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 628–635. https://doi.org/10.1145/3027063.3027072

Dewey, J. (1933). *How We Think*. Prometheus Books.

Diaz, R. (2024). Neural Resonator VST: Generate and Use Filters based on Arbitrary 2D Shapes and Materials. https://github.com/rodrigodzf/NeuralResonatorVST

Diaz, R., Hayes, B., Saitis, C., Fazekas, G., & Sandler, M. (2022). Rigid-Body Sound Synthesis with Differentiable Modal Resonators. *arXiv preprint*. https://doi.org/10.48550/arXiv.2210.15306

Dijk, E. T. K.-v., Westerink, J. H. D. M., Beute, F., & IJsselsteijn, W. A. (2017). Personal Informatics, Self-Insight, and Behavior Change: A Critical Review of Current Literature. *Human–Computer Interaction*, *32*(5-6), 268–296. https://doi.org/10.1080/07370024.2016.1276456

Dix, A. (2020). *Statistics for HCI: Making Sense of Quantitative Data* (1st ed.). Springer Cham. https://doi.org/https://doi.org/10.1007/978-3-031-02228-9

Doherty, K., & Doherty, G. (2018). Engagement in HCI: Conception, Theory and Measurement. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3234149

Downie, J. S., Byrd, D., & Crawford, T. (2009). Ten Years of ISMIR: Reflections on Challenges and Opportunities. *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009*, 13–18. https://archives.ismir.net/ismir2009/invited/000000.pdf

Eck, D., & Schmidhuber, J. (2002). *A First Look at Music Composition Using LSTM Recurrent Neural Networks*. Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale.

Edmonds, E. (2022). AI, Creativity, and Art. In C. Vear & F. Poltronieri (Eds.), *The Language of Creative AI: Practices, Aesthetics and Structures* (pp. 57–71). Springer International Publishing. https://doi.org/10.1007/978-3-031-10960-7_4

Edmonds, E., Muller, L., & Connell, M. (2006). On Creative Engagement. *Visual Communication*, *5*(3), 307–322. https://doi.org/10.1177/1470357206068461

Edmonds, E. A. (2014). Human Computer Interaction, Art and Experience. In L. Candy & S. Ferguson (Eds.), *Interactive Experience in the Digital Age: Evaluating New Art Practice* (pp. 11–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-04510-8_2

Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: An Overview. *Historical Social Research / Historische Sozialforschung*, *36*(4 (138)), 273–290. http://www.jstor.org/stable/23032294

Ens, J., & Pasquier, P. (2020). MMM : Exploring Conditional Multi-Track Music Generation with the Transformer. *arXiv preprint*. https://doi.org/10.48550/arXiv.2008.06048

Fabius, O., & Van Amersfoort, J. R. (2014). Variational Recurrent Auto-Encoders. *arXiv preprint*. https://doi.org/10.48550/arXiv.1412.6581

Fails, J. A., & Olsen, D. R. (2003). Interactive Machine Learning. *Proceedings of the 8th International Conference on Intelligent User Interfaces*, 39–45. https://doi.org/10.1145/604045.604056

Farbood, M. M., Pasztor, E., & Jennings, K. (2004). Hyperscore: A Graphical Sketchpad for Novice Composers. *IEEE Computer Graphics and Applications*, *24*(1), 50–54. https://doi.org/10.1109/MCG.2004.1255809

Fdili Alaoui, S. (2019). Making an Interactive Dance Piece: Tensions in Integrating Technology in Art. *Proceedings of the 2019 on Designing Interactive Systems Conference*, 1195–1208. https://doi.org/10.1145/3322276.3322289

Fdili Alaoui, S. (2023). *Dance-Led Research* [PhD Thesis]. Université Paris Saclay (COMUE).

Fiebrink, R., Cook, P. R., & Trueman, D. (2011). Human Model Evaluation in Interactive Supervised Learning. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 147–156. https://doi.org/10.1145/1978942.1978965

Fiebrink, R., & Sonami, L. (2020). Reflections on Eight Years of Instrument Creation with Machine Learning. In R. Michon & F. Schroeder (Eds.), *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)* (pp. 237–242). Birmingham City University. https://doi.org/10.5281/zenodo.4813334

Fiebrink, R., Trueman, D., Britt, C., Nagai, M., Kaczmarek, K., Early, M., Daniel, M., Hege, A., & Cook, P. (2012). Toward Understanding Human-Computer Interaction in Composing the Instrument. *Proceedings of the International Computer Music Conference (ICMC).*

Finney, S. J., & DiStefano, C. (2006). Non-normal and Categorical Data in Structural Equation Modeling. *Structural Equation Modeling: A Second Course*, *10*(6), 269–314.

Fischer, G. (2004). Social Creativity: Turning Barriers into Opportunities for Collaborative Design. *Proceedings of the Eighth Conference on Participatory Design: Artful Integration: Interweaving Media, Materials and Practices - Volume 1*, 152–161. https://doi.org/10.1145/1011870.1011889

Fleck, R., & Fitzpatrick, G. (2010). Reflecting on Reflection: Framing a Design Landscape. *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*, 216–223. https://doi.org/10.1145/1952222.1952269

Folkestad, G., Hargreaves, D. J., & Lindström, B. (1998). Compositional Strategies in Computer-Based Music-Making. *British Journal of Music Education*, *15*(1), 83–97. https://doi.org/10.1017/S0265051700003788

Ford, C., & Bryan-Kinns, N. (2022a). Identifying Engagement in Children's Interaction Whilst Composing Digital Music at Home. *Proceedings of the Fourteenth ACM Conference on Creativity and Cognition*, 443–456. https://doi.org/10.1145/3527927.3532794

Ford, C., & Bryan-Kinns, N. (2022b). Speculating on Reflection and People's Music Co-Creation with AI. *Workshop on Generative AI and HCI at the CHI Conference on Human Factors in Computing Systems 2022.* https://qmro.qmul.ac.uk/xmlui/handle/123456789/80144

Ford, C., Bryan-Kinns, N., & Nash, C. (2021). Creativity in Children's Digital Music Composition. In R. Dannenberg & X. Xiao (Eds.), *Proceedings of New Interfaces for Musical Expression (NIME) 2021.* https://nime.pubpub.org/pub/ker5w948/

Ford, C., & Nash, C. (2020, July). An Iterative Design 'by proxy' Method for Developing Educational Music Interfaces. In R. Michon & F. Schroeder (Eds.), *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME) 2020* (pp. 279–284). Birmingham City University. https://doi.org/10.5281/zenodo.4813361

Freeman, W. J. (1998). A Neurobiological Role of Music in Social Bonding. In B. M. N Wallin & S. Brown (Eds.), *The Origins of Music* (pp. 411–424). MIT Press. https://escholarship.org/uc/item/9025x8rt

Frich, J., MacDonald Vermeulen, L., Remy, C., Biskjaer, M. M., & Dalsgaard, P. (2019). Mapping the Landscape of Creativity Support Tools in HCI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3290605.3300619

Gamboa, M., Heron, M. J., Sturdee, M., & Belford, P. H. (2023). Screenshots as Photography in Gamescapes: An Annotated Psychogeography of Imaginary Places. *Proceedings of the 15th Conference on Creativity and Cognition*, 506–518. https://doi.org/10.1145/3591196.3593370

Garibay, O. O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., Havens, J. C., Jirotka, M., Kacorri, H., Karwowski, W., Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., . . . Xu, W. (2023). Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of Human–Computer Interaction*, *39*(3), 391–437. https://doi.org/10.1080/10447318.2022.2153320

Gaver, W. (2012). What Should We Expect from Research through Design? [event-place: Austin, Texas, USA]. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 937–946. https://doi.org/10.1145/2207676.2208538

Gaver, W. W., Beaver, J., & Benford, S. (2003). Ambiguity as a Resource for Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 233–240. https://doi.org/10.1145/642611.642653

Gioti, A.-M., Einbond, A., & Born, G. (2022). Composing the Assemblage: Probing Aesthetic and Technical Dimensions of Artistic Creation with Machine Learning. *Computer Music Journal*, *46*(4), 62–80. https://doi.org/10.1162/comj_a_00658

Glăveanu, V. P., & Kaufman, J. C. (2021). Creativity: A Historical Perspective. In J. C. Kaufman & R. J. Sternberg (Eds.), *Creativity: An Introduction* (pp. 1–16). Cambridge University Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, *27*.

Grant, A. M., Franklin, J., & Langford, P. (2002). The Self-Reflection and Insight Scale: A New Measure of Private Self-consciousness. *Social Behavior and Personality: An International Journal*, *30*(8), 821–836. https://doi.org/10.2224/sbp.2002.30.8.821

Guckelsberger, C., Salge, C., & Colton, S. (2017). Addressing the "Why?" in Computational Creativity: A Non-Anthropocentric, Minimal Model of Intentional Creative Agency. *Proceedings of the 8th International Conference on Computational Creativity*.

Guilford, J. (1950). Creativity. *American Psychologist*, *5*(9), 444–454.

Guillaumier, C. (2016). Reflection as Creative Process: Perspectives, Challenges and Practice. *Arts and Humanities in Higher Education*, *15*(3-4), 353–363. https://doi.org/10.1177/1474022216647381

Habermas, J. (1987). *Knowledge and Human Interests.* Polity Press.

Hadjeres, G., Pachet, F., & Nielsen, F. (2017). DeepBach: A Steerable Model for Bach Chorales Generation. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 1362–1371.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate Data Analysis (4th Ed.): With Readings*. Prentice-Hall, Inc.

Hallnäs, L., & Redström, J. (2001). Slow Technology – Designing for Reflection. *Personal and Ubiquitous Computing*, *5*(3), 201–212. https://doi.org/10.1007/PL00000019

Hamilton, A. (2020). The Aesthetics of Imperfection Reconceived: Improvisations, Compositions, and Mistakes. *The Journal of Aesthetics and Art Criticism*, *78*(3), 289–302. https://doi.org/10.1111/jaac.12749

Harley, J. (2002). The Electroacoustic Music of Iannis Xenakis. *Computer Music Journal*, *26*(1), 33–57. http://www.jstor.org/stable/3681399

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hartson, H. R., & Hix, D. (1989). Human-Computer Interface Development: Concepts and Systems for its Management. *ACM Computing Surveys*, *21*(1), 5–92. https://doi.org/10.1145/62029.62031

Hazzard, A., Greenhalgh, C., Kallionpaa, M., Benford, S., Veinberg, A., Kanga, Z., & McPherson, A. (2019). Failing with Style: Designing for Aesthetic Failure in Interactive Performance. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300260

Hepburn, A., & Bolden, G. B. (2012, November). The Conversation Analytic Approach to Transcription. In *The Handbook of Conversation Analysis* (pp. 57–76). https://doi.org/10.1002/9781118325001.ch4

Herremans, D., Chuan, C.-H., & Chew, E. (2017). A Functional Taxonomy of Music Generation Systems. *ACM Computing Surveys*, *50*(5). https://doi.org/10.1145/3108242

Hertzmann, A. (2018). Can Computers Create Art? *Arts*, *7*(2). https://doi.org/10.3390/arts7020018

Hewett, T. T. (2005). Informing the Design of Computer-based Environments to Support Creativity. *International Journal of Human-Computer Studies*, *63*(4), 383–409. https://doi.org/https://doi.org/10.1016/j.ijhcs.2005.04.004

Hewett, T., Czerwinski, M., Terry, M., Nunamaker, J., Candy, L., Kules, B., & Sylvan, E. (2005). Creativity Support Tool Evaluation Methods and Metrics. *National Science Foundation (NSF) Workshop Report on Creativity Support Tools*, 10–24.

Hiller, L., & Isaacson, L. (1957). *Illiac Suite: For String Quartet*. New Music Edition.

Höök, K., Caramiaux, B., Erkut, C., Forlizzi, J., Hajinejad, N., Haller, M., Hummels, C. C., Isbister, K., Jonsson, M., Khut, G., et al. (2018). Embracing First-person Perspectives in Soma-based Design. *Informatics*, *5*, 8.

Hoque, M. N., Ghai, B., & Elmqvist, N. (2022). DramatVis Personae: Visual Text Analytics for Identifying Social Biases in Creative Writing. *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 1260–1276. https://doi.org/10.1145/3532106.3533526

Huang, C. Z. A., & Chew, E. (2005). Palestrina Pal: A Grammar Checker for Music Compositions in the Style of Palestrina. *Proceedings of the 5th Conference on Understanding and Creating Music.*

Huang, C.-Z. A., Koops, H. V., Newton-Rex, E., Dinculescu, M., & Cai, C. J. (2020). AI Song Contest: Human-AI Co-Creation in Songwriting. *21st International Society for Music Information Retrieval Conference.* https://arxiv.org/pdf/2010.05388.pdf

Hubbard, L. J., Adricula, N., Brown, C., Perkoff, E. M., Dudy, S., Colunga, E., & Yeh, T. (2023). The Dimensions of Reflection Coding Scheme: A New Tool for Measuring the Impact of Designing for Reflection in Early Childhood. *Proceedings of the 15th Conference on Creativity and Cognition*, 519–528. https://doi.org/10.1145/3591196.3593512

Hubbard, L. J., Chen, Y., Colunga, E., Kim, P., & Yeh, T. (2021). Child-Robot Interaction to Integrate Reflective Storytelling Into Creative Play. *Proceedings of the 13th Conference on Creativity and Cognition.* https://doi.org/10.1145/3450741.3465254

Hunt, S. (2021). *Empirical Studies in End-User Computer-Generated Music Composition Systems* [PhD Thesis]. University of the West of England. https://uwe-repository.worktribe.com/output/7239594

Hunt, S. J., Mitchell, T., & Nash, C. (2020, July). Composing Computer Generated Music, An Observational Study using IGME: The Interactive Generative Music Environment. In R. Michon & F. Schroeder (Eds.), *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 61–66). Birmingham City University. https://doi.org/10.5281/zenodo.4813222

Jackson, S. A., Martin, A. J., & Eklund, R. C. (2008). Long and Short Measures of Flow: The Construct Validity of the FSS-2, DFS-2, and New Brief Counterparts. *Journal of Sport and Exercise Psychology*, *30*(5), 561–587. https://doi.org/10.1123/jsep.30.5.561

Jeon, Y., Jin, S., Shih, P. C., & Han, K. (2021). FashionQ: An AI-driven Creativity Support Tool for Facilitating Ideation in Fashion Design. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3411764.3445093

Jourdan, T., & Caramiaux, B. (2023). Machine Learning for Musical Expression: A Systematic Literature Review. *Proceedings of the International Conference on New Interfaces for Musical Expression.* https://www.nime.org/proceedings/2023/nime2023_46.pdf

Kahneman, D. (2011). *Thinking, Fast and Slow.* Macmillan.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kaiser, H. F. (1970). A Second Generation Little Jiffy. *Psychometrika*, *35*(4), 401–415. https://doi.org/10.1007/BF02291817

Kaptein, M. (2016). Using Generalized Linear (Mixed) Models in HCI. In J. Robertson & M. Kaptein (Eds.), *Modern Statistical Methods for HCI* (pp. 251–274). Springer International Publishing. https://doi.org/10.1007/978-3-319-26633-6%5C_11

Karpathy, A. (2015). The Unreasonable Effectiveness of Recurrent Neural Networks. http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Kaschub, M., & Smith, J. (2009). *Minds on Music: Composition for Creative and Critical Thinking.* R&L Education.

Kember, D., Leung, D. Y. P., Jones, A., Loke, A. Y., McKay, J., Sinclair, K., Tse, H., Webb, C., Wong, F. K. Y., Wong, M., & Yeung, E. (2000). Development of a Questionnaire to Measure the Level of Reflective Thinking. *Assessment & Evaluation in Higher Education*, *25*(4), 381–395. https://doi.org/10.1080/713611442

Kennedy, M. A. (2002). Listening to the Music: Compositional Processes of High School Composers. *Journal of Research in Music Education*, *50*(2), 94–110. https://doi.org/10.2307/3345815

Kerne, A., Webb, A. M., Latulipe, C., Carroll, E., Drucker, S. M., Candy, L., & Höök, K. (2013). Evaluation Methods for Creativity Support Environments. *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, 3295–3298. https://doi.org/10.1145/2468356.2479670

Kerne, A., Webb, A. M., Smith, S. M., Linder, R., Lupfer, N., Qu, Y., Moeller, J., & Damaraju, S. (2014). Using Metrics of Curation to Evaluate Information-Based Ideation. *ACM Transactions on Computer-Human Interaction*, *21*(3). https://doi.org/10.1145/2591677

Kim, J., & Lerch, F. J. (1997). Why Is Programming (Sometimes) So Difficult? Programming as Scientific Discovery in Multiple Problem Spaces. *Information Systems Research*, *8*(1), 25–50. https://doi.org/10.1287/isre.8.1.25

Kim, J., Agrawala, M., & Bernstein, M. S. (2017). Mosaic: Designing Online Creative Communities for Sharing Works-in-Progress. *Proceedings of the 2017 ACM Con-*

ference on Computer Supported Cooperative Work and Social Computing, 246–258. https://doi.org/10.1145/2998181.2998195

Kim, J., Bagla, A., & Bernstein, M. S. (2015). Designing Creativity Support Tools for Failure. *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, 157–160. https://doi.org/10.1145/2757226.2764542

King, P. M., & Kitchener, K. S. (1994). Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults. In *Jossey-Bass Higher and Adult Education Series and Jossey-Bass Social and Behavioral Science Series.* ERIC Institute of Educational Sciences.

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv preprint.* https://doi.org/10.48550/ARXIV.1312.6114

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling.* Guilford Publications.

Kolb, D. A. (1984). *Experiential Learning: Experience as the Source of Learning and Development.* Englewood Cliffs, New Jersey: Prentice-Hall.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kreminski, M., & Mateas, M. (2021). Reflective Creators. *Proceedings of the Twelfth International Conference on Computational Creativity (ICCC 2021).* https://mkremins.github.io/publications/ReflectiveCreators_ICCC2021.pdf

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and Evaluation of a User Experience Questionnaire. In A. Holzinger (Ed.), *Proceedings of the Symposium on HCI and Usability for Education and Work* (pp. 63–76). Springer.

Levine, T. C. (2014). The Use of Blogging in Tertiary Healthcare Educational Settings to Enhance Reflective Learning in Nursing Leadership. *Journal for Nurses in Professional Development*, *30*(6). https://doi.org/10.1097/NND.0000000000000103

Lewis, J. R., & Erdinç, O. (2017). User Experience Rating Scales with 7, 11, or 101 Points: Does it Matter? *Journal of Usability Studies*, *12*(2), 73–91.

Lewis, M. (2023). AIxArtist: A First-person Tale of Interacting with Artificial Intelligence to Escape Creative Block. *Proceedings of the 1st International Workshop on Explainable AI for the Arts (XAIxArts), ACM Creativity and Cognition (C&C) 2023.* https://arxiv.org/abs/2308.11424

Lewis, M. (2025). Art, Identity, and AI: Navigating Authenticity in Creative Practice. *Proceedings of the 2025 Conference on Creativity and Cognition*, 916–930. https://doi.org/10.1145/3698061.3726959

Lewis, M., Sturdee, M., Gamboa, M., & Lengyel, D. (2023). Doodle Away: An Autoethnographic Exploration of Doodling as a Strategy for Self-Control Strength in Online Spaces. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544549.3582747

Li, Y., Zhang, H., & O'Rourke, E. (2024). The Undervalued Disciplinary and Emotional Support Provided By Teaching Assistants in Introductory Computer Science Courses. In R. Lindgren, T. I. Asino, E. A. Kyza, C. K. Looi, D. T. Keifert, & E. Suárez (Eds.), *Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024* (pp. 1498–1501). International Society of the Learning Sciences. https://repository.isls.org//handle/1/10735

Li, Y., Chen, M., Hunt, A., Zhang, H., & O'Rourke, E. (2024). Exploring the Interplay of Metacognition, Affect, and Behaviors in an Introductory Computer Science Course for Non-Majors [event-place: Melbourne, VIC, Australia]. *Proceedings of the 2024 ACM Conference on International Computing Education Research - Volume 1*, 27–41. https://doi.org/10.1145/3632620.3671119

Li, Y., Nwogu, J., Buddemeyer, A., Solyst, J., Lee, J., Walker, E., Ogan, A., & Stewart, A. E. (2023). "I Want to Be Unique From Other Robots": Positioning Girls as Co-Creators of Social Robots in Culturally-Responsive Computing Education. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581272

Liu, S. (2017). On the Semantic and Non-semantic Nature of Music. *Proceedings of the 4th International Conference on Education, Management, Arts, Economics and Social Science (ICEMAESS 2017)*, 297–300. https://doi.org/10.2991/icemaess-17.2017.66

Lloret-Segura, S., Ferreres-Traver, A., Hernandez-Baeza, A., & Tomas-Marco, I. (2014). Exploratory Item Factor Analysis: A Practical Guide Revised and Updated. *Anales de Psicología*, *30*(3), 1151–1169.

Löbbers, S., Barthet, M., & Fazekas, G. (2021). Sketching Sounds: An Exploratory Study on Sound-shape Associations. *Proceedings of the International Computer Music Conference 2021*, 275–280.

Locher, P. J. (2010). How Does a Visual Artist Create an Artwork? In *The Cambridge Handbook of Creativity* (pp. 131–144). Cambridge University Press. https://doi.org/10.1017/CBO9780511763205.010

Loth, J., Sarmento, P., Carr, C. J., Zukowski, Z., & Barthet, M. (2023). ProgGP: From GuitarPro Tablature Neural Generation to Progressive Metal Production. *Proceedings of the 16th International Symposium on Computer Music Multidisciplinary Research*. https://arxiv.org/abs/2307.05328

Louie, R., Coenen, A., Huang, C. Z., Terry, M., & Cai, C. J. (2020). Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376739

Louie, R., Engel, J., & Huang, C.-Z. A. (2022). Expressive Communication: Evaluating Developments in Generative Models and Steering Interfaces for Music Creation. *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, 405–417. https://doi.org/10.1145/3490099.3511159

Lovelace, A. A. (1843). Notes by August Ada Lovelace. *Taylor's Scientific Memoirs*, *3*, 666–731.

Loy, G. (1985). Musicians Make a Standard: The MIDI Phenomenon [Publisher: The MIT Press]. *Computer Music Journal*, *9*(4), 8–26. Retrieved July 6, 2025, from http://www.jstor.org/stable/3679619

Lubart, T. (2005). How Can Computers Be Partners in the Creative Process: Classification and Commentary on the Special Issue. *International Journal of Human-Computer Studies*, *63*(4), 365–369. https://doi.org/10.1016/j.ijhcs.2005.04.002

Lubart, T. I. (2001). Models of the Creative Process: Past, Present and Future. *Creativity Research Journal*, *13*(3-4), 295–308. https://doi.org/10.1207/S15326934CRJ1334_07

Lucero, A. (2018). Living Without a Mobile Phone: An Autoethnography. *Proceedings of the 2018 Designing Interactive Systems Conference*, 765–776. https://doi.org/10.1145/3196709.3196731

Lucero, A., Desjardins, A., Neustaedter, C., Höök, K., Hassenzahl, M., & Cecchinato, M. E. (2019). A Sample of One: First-Person Research Methods in HCI. *Companion Publication of the 2019 Designing Interactive Systems Conference*, 385–388. https://doi.org/10.1145/3301019.3319996

Magnusson, T. (2019). *Sonic Writing: Technologies of Material, Symbolic, and Signal Inscriptions*. Bloomsbury Publishing USA.

Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and Validation of the Multimodal Presence Scale for Virtual Reality Environments: A Confirmatory Factor Analysis and Item Response Theory Approach. *Computers in Human Behavior*, *72*, 276–285. https://doi.org/10.1016/j.chb.2017.02.066

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Manovich, L. (2011). Inside Photoshop. *Computational Culture*, *1*. http://computationalculture.net/inside-photoshop

Matsunaga, M. (2010). How to Factor-analyze your Data Right: Do's, Don'ts, and How-to's. *International Journal of Psychological Research*, *3*(1), 97–110. https://doi.org/10.21500/20112084.854

McAdams, S. (2004). Problem-Solving Strategies in Music Composition: A Case Study. *Music Perception*, *21*(3), 391–429. https://doi.org/10.1525/mp.2004.21.3.391

McCarthy, L., Reas, C., & Fry, B. (2015). *Getting Started with P5.js: Making Interactive Graphics in JavaScript and Processing*. Maker Media Inc.

McGarry, G., Tolmie, P., Benford, S., Greenhalgh, C., & Chamberlain, A. (2017). "They're All Going out to Something Weird": Workflow, Legacy and Metadata in the Music Production Process. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 995–1008. https://doi.org/10.1145/2998181.2998325

McKay, C., & Fujinaga, I. (2006). jSymbolic: A feature extractor for MIDI files. *Proceedings of the International Computer Music Conference*, 302–5.

McLean, A., & Wiggins, G. (2010). Tidal–Pattern Language for the Live Coding of Music. *Proceedings of the 7th Sound and Music Computing Conference*, 331–334.

Medley, S., Zaman, B., & Haimes, P. (2020). The Role of Cuteness Aesthetics in Interaction. In R. Rousi, J. Leikas, & P. Saariluoma (Eds.), *Emotions in Technology Design: From Experience to Ethics* (pp. 125–138). Springer International Publishing. https://doi.org/10.1007/978-3-030-53483-7_8

Millen, D. R. (2000). Rapid Ethnography: Time Deepening Strategies for HCI Field Research. *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, 280–286. https://doi.org/10.1145/347642.347763

Mols, I., van den Hoven, E., & Eggen, B. (2016). Informing Design for Reflection: An Overview of Current Everyday Practices. *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. https://doi.org/10.1145/2971485.2971494

Montero, B. G. (2016). *Thought in Action: Expertise and the Conscious Mind*. Oxford University Press.

Moon, J. A. (2013). *Reflection in Learning and Professional Development: Theory and Practice*. Routledge.

Morrison, L., & McPherson, A. (2024). Entangling Entanglement: A Diffractive Dialogue on HCI and Musical Interactions. *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642171

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLOS ONE*, *9*(2), 1–23. https://doi.org/10.1371/journal.pone.0089642

Muller, M., Weisz, J. D., & Geyer, W. (2020). Mixed Initiative Generative AI Interfaces: An Analytic Framework for Generative AI Applications. *Proceedings of the 'The Future of Co-Creative Systems-A Workshop on Human-Computer Co-Creativity" at the 11th International Conference on Computational Creativity (ICCC 2020).*

Müller, H., Sedley, A., & Ferrall-Nunge, E. (2014). Survey Research in HCI. In J. S. Olson & W. A. Kellogg (Eds.), *Ways of Knowing in HCI* (pp. 229–266). Springer New York. https://doi.org/10.1007/978-1-4939-0378-8%5C_10

Murray-Browne, T., & Tigas, P. (2021). Latent Mappings: Generating Open-Ended Expressive Mappings Using Variational Autoencoders. *Proceedings of New Interfaces for Musical Expression (NIME).* https://doi.org/10.21428/92fbeb44.9d4bcd4b

Nakamura, J., & Csíkszentmihályi, M. (2009). Flow Theory and Research. In C. R. Snyder, S. J. Lopez, L. M. Edwards, & S. C. Marques (Eds.), *Handbook of Positive Psychology* (pp. 195–206). Oxford University Press.

Nash, C. (2011). *Supporting Virtuosity and Flow in Computer Music* [PhD Thesis]. University of Cambridge.

Nash, C. (2014). Manhattan: End-User Programming for Music. In K. Tahiroğlu, R. Fiebrink, A. Tana, & B. Caramiaux (Eds.), *Proceedings of the International Conference on New Interfaces for Musical Expression* (pp. 221–226). Goldsmiths. https://www.nime.org/proceedings/2014/nime2014%5C_371.pdf

Nash, C., & Blackwell, A. (2012). Liveness and Flow in Notation Use. In B. Gillespie (Ed.), *Proceedings of the International Conference on New Interfaces for Musical Expression 2012*. University of Michigan. https://www.nime.org/proceedings/2012/nime2012%5C_217.pdf

Nelson, B., & Rawlings, D. (2009). How Does It Feel? The Development of the Experience of Creativity Questionnaire. *Creativity Research Journal*, *21*(1), 43–53. https://doi.org/10.1080/10400410802633442

Nicholas, M. J., Sterman, S., & Paulos, E. (2022). Creative and Motivational Strategies Used by Expert Creative Practitioners. *Proceedings of the 14th Conference on Creativity and Cognition*, 323–335. https://doi.org/10.1145/3527927.3532870

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. *Workshop on Making Sense of Microposts: Big Things Come in Small Packages*, 93–98. http://arxiv.org/abs/1103.2903

Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann.

Nilsson, B., & Folkestad, G. (2005). Children's Practice of Computer-based Composition. *Music Education Research*, *7*(1), 21–37. https://doi.org/10.1080/14613800500042042

Noel-Hirst, A., & Bryan-Kinns, N. (2023). An Autoethnographic Exploration of XAI in Algorithmic Composition. *The 1st International Workshop on Explainable AI* *for the Arts (XAIxArts), ACM Creativity and Cognition (C&C) 2023*. https://arxiv.org/abs/2308.06089

Nordmoen, C., & McPherson, A. P. (2022). Making Space for Material Entanglements: A Diffractive Analysis of Woodwork and the Practice of Making an Interactive System. *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 415–423. https://doi.org/10.1145/3532106.3533572

Norman, D. A. (1993). *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Addison-Wesley Publishing Company.

O'Brien, H., & Toms, E. G. (2008). What is User Engagement? A Conceptual Framework for Defining User Engagement with Technology. *Journal of the American Society for Information Science & Technology*, *59*(6). http://dx.doi.org/10.14288/1.0107445

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human-Computer Studies*, *112*, 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

Oliveros, P. (2005). *Deep Listening: A Composer's Sound Practice*. IUniverse.

Ooi, S. M., Fisher, P., & Coker, S. (2021). A Systematic Review of Reflective Practice Questionnaires and Scales for Healthcare Professionals: A Narrative Synthesis. *Reflective Practice*, *22*(1), 1–15. https://doi.org/10.1080/14623943.2020.1801406

OpenAI. (2022). ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/

Oppenlaender, J., Milland, K., Visuri, A., Ipeirotis, P., & Hosio, S. (2020). Creativity on Paid Crowdsourcing Platforms. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376677

O'Reilly, S. L., & Milner, J. (2015). Transitions in Reflective Practice: Exploring Student Development and Preferred Methods of Engagement. *Nutrition & Dietetics*, *72*(2), 150–155. https://doi.org/10.1111/1747-0080.12134

Pachet, F. (2003). The Continuator: Musical Interaction With Style. *Journal of New Music Research*, *32*(3), 333–341. https://doi.org/10.1076/jnmr.32.3.333.16861

Pati, A., Lerch, A., & Hadjeres, G. (2019). Learning To Traverse Latent Spaces For Musical Score Inpainting. *Proceedings of the 20th International Society for Music Information Retrieval Conference*. https://archives.ismir.net/ismir2019/paper/000040.pdf

Pérez, R. P. Ý., & Sharples, M. (2001). MEXICA: A Computer Model of a Cognitive Account of Creative Writing. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(2), 119–139. https://doi.org/10.1080/09528130010029820

Phi, M. (2018). Illustrated Guide to LSTMs and GRUs: A Step by Step Explanation. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

Poupyrev, I., Lyons, M. J., Fels, S., & Blaine (Bean), T. (2001). New Interfaces for Musical Expression. *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, 491–492. https://doi.org/10.1145/634067.634348

Preece, J., Sharp, H., & Rogers, Y. (2011). *Interaction Design: Beyond Human-Computer Interaction* (Third). John Wiley & Sons.

Preston, C. C., & Colman, A. M. (2000). Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. *Acta Psychologica*, *104*(1), 1–15. https://www.sciencedirect.com/science/article/pii/S0001691899000505

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, *1*(8), 9.

Rajcic, N., Llano Rodriguez, M. T., & McCormack, J. (2024). Towards a Diffractive Analysis of Prompt-Based Generative AI. *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3641971

Reicherts, L., Park, G. W., & Rogers, Y. (2022). Extending Chatbots to Probe Users: Enhancing Complex Decision-Making Through Probing Conversations. *Proceedings of the 4th Conference on Conversational User Interfaces*. https://doi.org/10.1145/3543829.3543832

Reimer, P. J. C., & Wanderley, M. M. (2021). Embracing Less Common Evaluation Strategies for Studying User Experience in NIME. *Proceedings of the International Conference on New Interfaces for Musical Expression*. https://doi.org/10.21428/92fbeb44.807a000f

Remy, C., MacDonald Vermeulen, L., Frich, J., Biskjaer, M. M., & Dalsgaard, P. (2020). Evaluating Creativity Support Tools in HCI Research. *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 457–476. https://doi.org/10.1145/3357236.3395474

Renner, B., Kimmerle, J., Cavael, D., Ziegler, V., Reinmann, L., & Cress, U. (2014). Web-Based Apps for Reflection: A Longitudinal Study With Hospital Staff. *Journal of Medical Internet Research*, *16*(3), 85. https://doi.org/10.2196/jmir.3040

Resnick, M., Myers, B., Nakakoji, K., Shneiderman, B., Pausch, R., Selker, T., & Eisenberg, M. (2005). Design Principles for Tools to Support Creative Thinking. *National Science Foundation (NSF) Workshop Report on Creativity Support Tools*, 25–36.

Rezwana, J., & Maher, M. L. (2022). Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM*

Transactions on Computer-Human Interaction, *30*(5), 1–28. https://doi.org/10.1145/3519026

Rhodes, M. (1961). An Analysis of Creativity. *The Phi Delta Kappan*, *42*(7), 305–310. http://www.jstor.org/stable/20342603

Richards, R. (2010). Everyday Creativity: Process and Way of Life – Four Key Issues. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge Handbook of Creativity* (pp. 189–215). Cambridge University Press.

Rivard, K., & Faste, H. (2012). How Learning Works in Design Education: Educating for Creative Awareness through Formative Reflexivity. *Proceedings of the Designing Interactive Systems Conference*, 298–307. https://doi.org/10.1145/2317956.2318002

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., & Eck, D. (2018). A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. *Proceedings of the 35th International Conference on Machine Learning (ICML '18)*, *80*. https://proceedings.mlr.press/v80/roberts18a/roberts18a.pdf

Roberts, A., Hawthorne, C., & Simon, I. (2018). Magenta.js: A JavaScript API for Augmenting Creativity with Deep Learning. *Joint Workshop on Machine Learning for Music (ICML)*.

Robson, N., McPherson, A., & Bryan-Kinns, N. (2024). Thinking with Sound: Exploring the Experience of Listening to an Ultrasonic Art Installation. *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642616

Roldan, W., Li, Z., Gao, X., Kay Strickler, S., Marie Hishikawa, A., E. Froehlich, J., & Yip, J. (2021). Pedagogical Strategies for Reflection in Project-based HCI Education with End Users. *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, 1846–1860. https://doi.org/10.1145/3461778.3462113

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rossmy, B. (2022). Buttons, Sliders, and Keys – A Survey on Musical Grid Interface Standards. *Proceedings of the International Conference on New Interfaces for Musical Expression*. https://doi.org/10.21428%2F92fbeb44.563bfea9

Saitis, C., Sette, B. M. D., Shier, J., Tian, H., Zheng, S., Skach, S., Reed, C. N., & Ford, C. (2024). Timbre Tools: Ethnographic Perspectives on Timbre and Sonic Cultures in Hackathon Designs. *Proceedings of the 2024 ACM International Audio Mostly Conference - Explorations in Sonic Cultures (AM '24)*, 16. https://doi.org/10.1145/3678299.3678322

Sanders, E. B.-N., & Stappers, P. J. (2008). Co-creation and the New Landscapes of Design. *CoDesign*, *4*(1), 5–18. https://doi.org/10.1080/15710880701875068

Sapp, D. D. (1992). The Point of Creative Frustration and the Creative Process: A New Look at an Old Model. *Journal of Creative Behavior*, *26*, 21–28. https://api.semanticscholar.org/CorpusID:144088459

Sarmento, P., Kumar, A., Carr, C. J., Zukowski, Z., Barthet, M., & Yang, Y. (2021). DadaGP: A Dataset of Tokenized GuitarPro Songs for Sequence Models. *Proceedings of the 22nd International Society for Music Information Retrieval Conference*. https://arxiv.org/abs/2107.14653

Sarmento, P., Kumar, A., Chen, Y.-H., Carr, C., Zukowski, Z., & Barthet, M. (2023). GTR-CTRL: Instrument and Genre Conditioning for Guitar-focused Music Generation with Transformers. *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, 260–275.

Sas, C., & Dix, A. (2009). Designing for Reflection on Experience. *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, 4741–4744. https://doi.org/10.1145/1520340.1520730

Sayer, T. (2015). Cognition and Improvisation: Some Implications for Live Coding. *Proceedings of the First International Conference on Live Coding*, 87–92. https://doi.org/10.5281/zenodo.19328

Schön, D. A. (1983). *The Reflective Practitioner: How Professionals Think in Action*. Basic Books Inc.

Schön, D. A. (1987). *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions*. Jossey-Bass.

Schrepp, M. (2020). On the Usage of Cronbach's Alpha to Measure Reliability of UX Scales. *Journal of Usability Studies*, *15*(4).

Seligman, M. E., & Csíkszentmihályi, M. (2014). Positive Psychology: An Introduction. In M. Csíkszentmihályi (Ed.), *Flow and the Foundations of Positive Psychology* (pp. 279–298). Springer.

Sengers, P., Boehner, K., David, S., & Kaye, J. (2005). Reflective Design. *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, 49–58. https://doi.org/10.1145/1094562.1094569

Sharmin, M., & Bailey, B. P. (2013). ReflectionSpace: An Interactive Visualization Tool for Supporting Reflection-on-Action in Design. *Proceedings of the 9th ACM Conference on Creativity & Cognition*, 83–92. https://doi.org/10.1145/2466627.2466645

Sharples, M. (1996). An Account of Writing as Creative Design. In C. M. Levy & S. Randell (Eds.), *The Science of Writing* (pp. 127–148). Erlbaum.

Sheldon, K. M., Prentice, M., & Halusic, M. (2015). The Experiential Incompatibility of Mindfulness and Flow Absorption. *Social Psychological and Personality Science*, *6*(3), 276–283. https://doi.org/10.1177/1948550614555028

Shneiderman, B. (2002). Creativity Support Tools. *Communications of the ACM*, *45*(10), 116–120. https://doi.org/10.1145/570907.570945

Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press.

Shneiderman, B., Fischer, G., Czerwinski, M., Resnick, M., Myers, B., Candy, L., Edmonds, E., Eisenberg, M., Giaccardi, E., Hewett, T. T., Jennings, P., Kules, B., Nakakoji, K., Nunamaker, J., Pausch, R., Selker, T., Sylvan, E., & Terry, M. (2006). Creativity Support Tools: Report From a U.S. National Science Foundation Sponsored Workshop. *International Journal of Human-Computer Interaction*, *20*(2), 61–77. https://doi.org/10.1207/s15327590ijhc2002%5C_1

Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). Where to Look: A Study of Human-Robot Engagement. *Proceedings of the 9th International Conference on Intelligent User Interfaces*, 78–84. https://doi.org/10.1145/964442.964458

Sloboda, J. (1985). *The Musical Mind: The Cognitive Psychology of Music*. Oxford University Press.

Slovák, P., Frauenberger, C., & Fitzpatrick, G. (2017). Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2696–2707. https://doi.org/10.1145/3025453.3025516

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72–101. http://www.jstor.org/stable/1412159

Spiel, K. (2021). “Why are they all obsessed with Gender?” — (Non)binary Navigations through Technological Infrastructures. *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, 478–494. https://doi.org/10.1145/3461778.3462033

Spoto, A., & Oleynik, N. (2018). Library of Mixed-Initiative Creative Interfaces. http://mici.codingconduct.cc/aboutmicis/

Stein, M. I. (1953). Creativity and Culture. *The Journal of Psychology*, *36*(2), 311–322. https://doi.org/10.1080/00223980.1953.9712897

Steinaker, N., & Bell, M. R. (1975). A Proposed Taxonomy of Educational Objectives: The Experiential Domain. *Educational Technology*, *15*(1), 14–16.

Sterman, S. (2022, August). *Process-Sensitive Creativity Support Tools* [PhD Thesis]. University of California, Berkeley. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-207.html

Sterman, S., Nicholas, M. J., Vivrekar, J., Mindel, J. R., & Paulos, E. (2023). Kaleidoscope: A Reflective Documentation Tool for a User Interface Design Course. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581255

Sturdee, M., Lewis, M., Strohmayer, A., Spiel, K., Koulidou, N., Alaoui, S. F., & Urban Davis, J. (2021). A Plurality of Practices: Artistic Narratives in HCI Research. *Proceedings of the 13th Conference on Creativity and Cognition*. https://doi.org/10.1145/3450741.3466771

Sturm, B. (2022). Generative AI Helps One Express Things for Which They May Not Have Expressions (Yet). *Workshop on Generative AI and HCI at the CHI Conference on Human Factors in Computing Systems 2022*. https://kth.diva-portal.org/smash/get/diva2:1757906/FULLTEXT01.pdf

Sturm, B. L., Ben-Tal, O., Monaghan, Ú., Collins, N., Herremans, D., Chew, E., Hadjeres, G., Deruty, E., & Pachet, F. (2019). Machine Learning Research that Matters for Music Creation: A Case Study. *Journal of New Music Research*, *48*(1), 36–55. https://doi.org/10.1080/09298215.2018.1515233

Sturm, B. L., Santos, J. F., Ben-Tal, O., & Korshunova, I. (2016). Music Transcription Modelling and Composition using Deep Learning. *arXiv preprint*. https://arxiv.org/abs/1604.08723

Suh, M., Youngblom, E., Terry, M., & Cai, C. J. (2021). AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445219

Swanwick, K., & Tillman, J. (1986). The Sequence of Musical Development: A Study of Children's Composition. *British Journal of Music Education*, *3*(3), 305–339. https://doi.org/10.1017/S0265051700000814

Taherdoost, H., Sahibuddin, S., & Jalaliyoon, N. (2014). Exploratory Factor Analysis; Concepts and Theory. In J. Balicki (Ed.), *Advances in Applied and Pure Mathematics* (pp. 375–382, Vol. 27). WSEAS. https://hal.archives-ouvertes.fr/hal-02557344

Tang, G., & Zhou, W. (2020). The Study on Self-consciousness in Flow. *Philosophy Study*, *10*(10), 615–621. https://doi.org/10.17265/2159-5313/2020.10.002

Tchemeube, R. B., Ens, J., Plut, C., Pasquier, P., Safi, M., Grabit, Y., & Rolland, J.-B. (2023). Evaluating Human-AI Interaction via Usability, User Experience and Acceptance Measures for MMM-C: A Creative AI System for Music Composition. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, 5769–5778. https://doi.org/10.24963/ijcai.2023/640

Tchemeube, R. B., Ens, J. J., & Pasquier, P. (2022). Calliope: A Co-Creative Interface for Multi-Track Music Generation. *Proceedings of the ACM Creativity and Cognition Conference*, 608–611. https://doi.org/10.1145/3527927.3535200

Thelle, N. J. W., & Fiebrink, R. (2022). How Do Musicians Experience Jamming With a Co-Creative "AI"? *Proceedings of the 36th Conference on Neural Information*

Processing Systems (NeurIPS 2022). https://ualresearchonline.arts.ac.uk/id/eprint/20204/1/ThelleFiebrink_NeurIPSCreativity2023.pdf

Thiebaut, J.-B., Healey, P. G. T., & Bryan-Kinns, N. (2008). Drawing Electroacoustic Music. *Proceedings of the International Computer Music Conference 2008.*

Tseng, T., & Resnick, M. (2016). Spin: Examining the Role of Engagement, Integration, and Modularity in Supporting Youth Creating Documentation. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, 996–1007. https://doi.org/10.1145/2901790.2901868

Unesco, I. (2020). Basic Texts of the 2003 Convention for the Safeguarding of the Intangible Cultural Heritage.

Vahlo, J., & Karhulahti, V.-M. (2020). Challenge Types in Gaming Validation of Video Game Challenge Inventory (CHA). *International Journal of Human-Computer Studies*, *143*, 102473. https://doi.org/10.1016/j.ijhcs.2020.102473

Vanka, S. S., Safi, M., Rolland, J.-B., & Fazekas, G. (2023). Adoption of AI Technology in the Music Mixing Workflow: An Investigation. *154th Audio Engineering Society Convention.* https://arxiv.org/pdf/2304.03407

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Wagener, N., Reicherts, L., Zargham, N., Bartłomiejczyk, N., Scott, A. E., Wang, K., Bentvelzen, M., Stefanidi, E., Mildner, T., Rogers, Y., & Niess, J. (2023). SelVReflect: A Guided VR Experience Fostering Reflection on Personal Challenges. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3544548.3580763

Wald, G. (2023). "A Queer Black Woman Invented Rock-and-roll": Rosetta Tharpe, Memes, and Memory Practices in the Digital Age [Publisher: Routledge]. *Feminist Media Studies*, *23*(3), 1075–1091. https://doi.org/10.1080/14680777.2020.1855224

Wallas, G. (1926). *The Art of Thought.* J. Cape.

Wang, Z., Chen, K., Jiang, J., Zhang, Y., Xu, M., Dai, S., & Xia, G. (2020). POP909: A Pop-song Dataset for Music Arrangement Generation. *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR).* Retrieved February 17, 2023, from https://zenodo.org/record/4245366

Webster, P. R. (2002). Creative Thinking in Music: Advancing a Model. In T. Sullivan & L. Willingham (Eds.), *Creativity and Music Education* (pp. 16–33). Canadian Music Educators Association.

Whittall, A. (2011). Composition. Oxford University Press. https://www.oxfordreference.com/view/10.1093/acref/9780199579037.001.0001/acref-9780199579037-e-1531

Wiggins, G. A. (2006). Searching for Computational Creativity. *New Generation Computing*, *24*(3), 209–222. https://doi.org/10.1007/BF03037332

Wilson, E., Fazekas, G., & Wiggins, G. (2023). On the Integration of Machine Agents into Live Coding. *Organised Sound*, *28*(2), 1–10. https://doi.org/10.1017/S1355771823000420

Wilson, E., Lawson, S., McLean, A., & Stewart, J. (2021). Autonomous Creation of Musical Pattern from Types and Models in Live Coding. *xCoAx 2021 9th Conference on Computation, Communication, Aesthetics & X*, 76–93. https://qmro.qmul.ac.uk/xmlui/handle/123456789/73475

Worrall, K., Yin, Z., & Collins, T. (2022). Comparative Evaluations in the Wild: Systems for the Expressive Rendering of Music. *IEEE Transactions on Artificial Intelligence*.

Worthington, R. L., & Whittaker, T. A. (2006). Scale Development Research: A Content Analysis and Recommendations for Best Practices. *The Counseling Psychologist*, *34*(6), 806–838.

Woźniak, P. W., Karolus, J., Lang, F., Eckerth, C., Schöning, J., Rogers, Y., & Niess, J. (2021). Creepy Technology: What Is It and How Do You Measure It? *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445299

Wu, Y. (2018). *Design and Evaluate Support for Non-musicians' Creative Engagement with Musical Interfaces* [PhD Thesis]. Queen Mary University of London.

Wu, Y., & Bryan-Kinns, N. (2019). Musicking with an Interactive Musical System: The Effects of Task Motivation and User Interface Mode on Non-musicians' Creative Engagement. *International Journal of Human-Computer Studies*, *122*, 61–77. https://doi.org/10.1016/j.ijhcs.2018.07.009

Xambó, A. (2022). Virtual Agents in Live Coding: A Review of Past, Present and Future Directions. *e-Contact!*, *21*(1). https://arxiv.org/abs/2106.14835

Xu, W. (2019). Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *Interactions*, *26*(4), 42–46. https://doi.org/10.1145/3328485

Yang, L.-C., & Lerch, A. (2020). On the Evaluation of Generative Models in Music. *Neural Computing and Applications*, *32*(9), 4773–4784. https://doi.org/10.1007/s00521-018-3849-7

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376301

Youngblood, G. (1998). Jean-Luc Godard: No Difference between Life and Cinema. In D. Sterritt (Ed.), *Jean-Luc Godard: Interviews* (pp. 9–49). University Press of Mississippi.

Younker, B. A. (2000). Thought Processes and Strategies of Students Engaged in Music Composition. *Research Studies in Music Education*, *14*(1), 24–39. https://doi.org/10.1177/1321103X0001400103

Yurman, P. (2021). Fluid Speculations: Drawing Artefacts in Watercolour as Experimentation in Research Through Design. *Proceedings of the Thirteenth ACM Conference on Creativity and Cognition*. https://doi.org/10.1145/3450741.3466777

Yurman, P., & Reddy, A. V. (2022). Drawing Conversations Mediated by AI [event-place: Venice, Italy]. *Proceedings of the 14th Conference on Creativity and Cognition*, 56–70. https://doi.org/10.1145/3527927.3531448

Zaki, M. J., & Meira Jr, W. (2020). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press.

Appendix

All appendix material can be found online at:

https://github.com/aim-qmul/Corey-Ford-PhD-Appendix.