The Next Generation of Implicit Brain-Computer Interfaces: Enhancing User Learning and Creativity

A dissertation

submitted by

Beste Filiz Yuksel, B.S., M.S.

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Computer Science

TUFTS UNIVERSITY

August 2016

Advisor: Robert J.K. Jacob

There is a fundamentally limited bandwidth of implicit and explicit communication between the human and computer. However, both humans and computers are complex machines, capable of sophisticated functions. If we could provide the computer with more implicit information about the human, without any additional effort on the part of the user, the computer could then respond more intelligently in return.

In this thesis, I expand upon the next generation of implicit, brain-computer interfaces by building two, real-time adaptive brain-computer interfaces (BCIs) based on users' cognitive workload in the previously unexplored fields of ***learning*** and ***creativity***. I demonstrate that users can learn with increased speed and accuracy with a BCI that guides learners' progress by measuring when they can cognitively handle more information and providing them with the next stage of learning at the right moment. I also build and evaluate an adaptive BCI that addresses the no-less tricky task of increasing user creativity in a musical task. Finally, I address the topic of whether users will actually *trust* such intelligent systems by building a trust evaluation system that reveals that building trust in human-computer interaction can be just as wiley as human-human interaction.

I suggest that measuring affective state in conjunction with cognitive state in the future would provide more accurate insight into the user's state, allowing for more personalized and intelligent adaptations by the next generation of implicit, brain-computer interfaces.

All glory to the teacher.

To my advisor, Rob Jacob, genuinely the sweetest person I have ever met and a leader in cutting-edge HCI research ever since there was HCI. Thank you for carrying out such exciting research, for guiding me through this most perilous journey, for moulding me as a researcher and as a better human. You are the best advisor anyone could have, I am so fortunate to have been under your wing. Like all your students before you, I will always be immeasurably grateful. Thank you for teaching me how to fly.

To Remco Chang, the unsung hero of this piece. On paper, his true collaboration, influence and input on my work is not reflected as he was officially neither my advisor nor co-advisor. Remco did not have to be the pillar of advice and support that he was throughout my PhD. But his influence on my work is immeasurable as his critical thinking and research skills have become intertwined into forming me as a researcher. The best brain-stormer I know, I will always be grateful for your vision and creativity, you have always pushed me to be better, and I am sure you will never stop, no matter how far I go.

To Mary Czerwinski, my female role-model and mentor at Microsoft Research who taught me so much in such a short space of time. I grew exponentially as a researcher and as a person after working with you. That was really the turning point for me on so many levels. I am so honored and privileged to know and be able to work with such a badass, groundbreaking, and intelligent woman such as you.

To my two labmates Evan Peck and Dan Afergan who guided, supported, helped, and cajoled me through my PhD, and made me laugh even during the hardest parts. Our work is so collaborative, and I could not have achieved what I did without all the many hours of work and brainstorming you both put in.

To my award-winning first advisee Kurt Oleson, you were my first experience in advising a student on a project, and it inspired me to continue with academia and work with other students. What a special intro into being an advisor; you were a star, and I was incredibly fortunate to be able to work with you.

To my two partners in musical crime: Garth Griffin and Kathleen Kuo, who understood my musical crises and always came to the rescue, I salute your knowledge!

To Erin Solovey, for her continued guidance and support in the world of fNIRS

and research!

To the members of the HCI and VALT Labs: Tomoki Shibata, Sam Hincks, Alvitta Ottley, Anzu Hakone, Ronna ten Brink, Fumeng Yang, to Jordan Crouser, for being a bastion of wisdom, to Eli Brown whose warmth, friendship and laughter was a joy throughout, to Lane Harrison, the king of stats. To Sergio Fanitini, Angelo Sassaroli, and Jana Kainerstorfer from the Biomedical Engineering Department and Paul Lehrman from Music Engineering. To Ben Hescott, teacher extraordinaire, I hope one day to be a fraction of the teacher you are. To Donna Cirelli and Jeannine Vangelist, the formidable Computer Science Admin team at Tufts University.

To the VIBE team at Microsoft Research including Asta Roseway, Kael Rowan, Paul Johns, Rob DeLine, Andrew Begel. To members of the Medical Devices group including Dan Morris and Jeremiah Wander.

To Penny Collisson from Team Six at Microsoft who introduced me to the importance of user trust in computers.

To two incredibly giving and loyal friends: Candice and Karen. I am not sure this would have happened without your support.

And to Denise, who I met at the end of this adventure, and then realized the very thought of her had been keeping me going from the beginning. Are you ready for the next adventure? :)

To my parents who worked so hard all their lives so that I could have a better life than they did. I hope this brings you some joy. Turns out I did end up as a doctor, afterall! Maybe you can stop worrying now :) But seriously, I was very fortunate to have been your daughter.

Finally, none of my achievements would have been possible without the exceptional guidance, friendship, and massive giving over the last 10 years of a most beloved teacher. This thesis is dedicated to you.

All glory to the teacher!

Contents

List of Figures **xii**

List of Tables **xviii**

Chapter 1 Introduction **1**

1.1 Motivation 1

1.1.1 Why Brain-Computer Interfaces? 2

1.2 Thesis Contributions 5

1.3 Thesis Overview 6

Chapter 2 Related Work **8**

2.1 Brain-Computer Interfaces 8

2.1.1 Traditional Approaches 8

2.1.2 Next Generation of Implicit Brain-Computer Interfaces . . . 13

2.1.3 Cognitive Workload and the Prefrontal Cortex 14

2.1.4 functional Near Infrared Spectroscopy (fNIRS) 16

2.2 Learning 18

2.2.1 Computer-Based Education 18

2.2.2 Finite Cognitive Workload and Learning 19

2.2.3 Measurement of Cognitive Workload in Learning Studies . . . 20

2.3 Creativity 26

2.3.1 Creativity in HCI 26

2.3.2 Musical Creativity and BCIs 27

2.4 Trust 29
2.4.1 Cognitive and Affective Trust 29
2.4.2 Trust between Humans and Computers 30
2.4.3 Physical Attractiveness Stereotype 32

Chapter 3 Experiment 1: Feasibility Study **35**

3.1 fNIRS System 35
3.1.1 Equipment and Measurements 35
3.1.2 Procedures 36
3.1.3 Data Analysis and Classification of fNIRS Signal 39
3.2 Experiment 1 Results and Discussion 45
3.3 Experiment 1 Conclusions 46

Chapter 4 Experiment 2: Learning **47**

4.1 Experiment 2 - Introduction 47
4.2 Experimental Design 49
4.3 Learning Task 50
4.3.1 Music for Learning Task 50
4.4 Algorithms Behind BACh 51
4.4.1 Initial Real Time Classification of Learning 51
4.4.2 BACh's Algorithm 53
4.5 Evaluation of Dependent Variables 55
4.5.1 Dependent Variables 55
4.5.2 Musical Assessment 58
4.6 Experiment 2 - Results 58
4.6.1 fNIRS Data 58
4.6.2 Performance Data 59
4.6.3 Questionnaire Data 62
4.6.4 Timings of Level Changes 63
4.7 Experiment 2 - Discussion 65
4.7.1 Modeling and Adapting to Cognitive Workload 66

4.7.2 Responding to Learners Individually 67
4.7.3 Expertise of Learners 68
4.8 Experiment 2 - Conclusion 68

Chapter 5 Experiment 3: Creativity **70**

5.1 Experiment 3 - Introduction 70
5.2 BrAAHMS: Iterative Designs 72
5.2.1 Design Iterations and Pilot Studies 72
5.3 Final Design for BrAAHMS 74
5.3.1 Musical Real-Time Adaptation 74
5.3.2 Sound Engine 75
5.3.3 Musical Guidance for Participants 76
5.4 Algorithm Behind BrAAHMS 77
5.4.1 Initial Real-Time Classification 77
5.4.2 BrAAHMS' Algorithm 78
5.5 Evaluation of BrAAHMS 78
5.5.1 Experimental Design 79
5.5.2 Participants 80
5.5.3 Experiment 3 - Results and Discussion 80
5.6 Experiment 3 - Discussion 86
5.7 Experiment 3 - Conclusion 87

Chapter 6 Experiment 4: Trust Evaluation System **89**

6.1 Experiment 4 - Introduction 89
6.2 Hypothesis 91
6.3 Method 92
6.3.1 Materials 92
6.3.2 Experimental Design 94
6.4 Mapping of Dependent Variables to Trust 97
6.4.1 Number of Questions With Button Presses 97
6.4.2 Self-reported Measures of Trust in Agent 98

6.4.3 Self-reported Measures of Perceived Accuracy of Agent 98
6.4.4 Number of Correct Suggestions Accepted 98
6.4.5 Number of Incorrect Suggestions Accepted 98
6.4.6 Length of Time Spent on Questions With Suggestions 99
6.4.7 Number of Timeouts With Suggestions 99
6.4.8 Number of Correct Answers 99
6.4.9 Number of Incorrect Answers 99
6.5 Results 100
6.5.1 Verification of Attractiveness Independent Variable 100
6.5.2 Dependent Variables 100
6.6 Experiment 4 - Discussion 105
6.6.1 Reliability Manipulation Check 105
6.6.2 Is Reliability More Important than Attractiveness in Building Trust in User-Agent Interactions? 106
6.6.3 Cognitive and Affective Trust 108
6.6.4 Physical versus Vocal Attractiveness in User-Agent Interactions 109
6.6.5 Limitations 110
6.7 Experiment 4 - Conclusion 111

Chapter 7 Conclusions **112**
7.1 Summary of Work and Contributions 112
7.2 Future Work 114
7.2.1 Future Work on Learning 114
7.2.2 Future Work on Creativity 116
7.2.3 Future Work on Trust 117
7.3 Closing Remarks 118

Bibliography **119**

Appendix A Musical Pieces Used During the fNIRS Training Task. **145**

Appendix B Questionnaires **158**

List of Figures

3.1 fNIRS equipment and experiment setup. Left: Subject is playing the piano keyboard while wearing the fNIRS sensor. Imagent is visible on the right. Right: An fNIRS sensor with light sources (left) and one detector (right). 36

3.2 Each fNIRS probe had 4 sources (five are shown but four were used) and one detector which were placed on the forehead to take measurements from the prefrontal cortex [161]. 36

3.3 N-back task: Each square is flashed on the screen one at a time. Left: 1-back elicits significantly lower cognitive workload than Right: 3-back which requires holding more information in the short-term memory. 37

3.4 An example of a hard piece, excerpt from Impromptu No.1, Opus 25 by Gabriel Fauré 38

3.5 An example of an easy piece 38

3.6 An example of a a) hard piece and b) easy piece. Both pieces are nine measures long. 38

3.7 Overall signal analysis and classification of the fNIRS system. 40

3.8 Relative change in optical intensity of one individual averaged across 30 trials for 30 seconds per trial for one channel. The two features, mean and linear regression slope, are highlighted. 41

3.9 Example of Top: Relative change in optical intensity of one individual averaged across 30 trials for 30 seconds per trial across all 16 channels. Bottom: The highlighted channels are the ones used to pick features by the LIBSVM machine learning algorithm. The algorithm picked different channels for different individuals. 43

3.10 Left: Mean change and standard error in oxy-Hb in Experiment 1 across all participants. Although each participant was modeled individually, the fNIRS signal exhibited a general trend with higher levels of oxy-Hb corresponding with hard pieces. Right: The mean change in oxy-Hb was significantly higher in participants when they played a hard piece than an easy piece ($p < .01$). 46

4.1 Scores used for BACh and Normal conditions. BACh increased the difficulty level by adding a full line of music when users' cognitive workload fell below a threshold, indicating that they could handle more information. 52

4.2 Examples of piano rolls from a participant's performance (top) compared with a computer generated groundtruth of the same piece (bottom). Incorrect notes are denoted by a cross, extra notes by a circle, and missed notes by a triangle. An error is a temporal group of notes and is denoted by an arrow pointing up at the column of notes containing an error. 57

4.3 Mean change and standard error in oxy-Hb across all participants. Although each participant was modeled individually, the fNIRS signal exhibited a general trend with higher levels of oxy-Hb corresponding with hard pieces. The mean change in oxy-Hb was significantly higher in participants when they played an hard piece than an easy piece ($p = .01$). 59

4.4 Slopegraphs showing the effects of BACh for each participant. Upward sloping lines (blue) are indicative of better performance with BACh. Downward sloping lines (green) indicate better performance in the normal condition. Horizontal lines that indicate equality are also indicated in green. Participants showed significantly better performance with BACh in all four conditions ($p < .01$). Interestingly, participants who were less skilled seemed to benefit more greatly from BACh. 61

4.5 Mean and standard error of participants' ratings of BACh and the normal condition. Participants showed noticeable preferences for BACh in how well they felt they mastered the piece, how correctly they felt they played, and how easy they felt the piece was to learn. 62

4.6 Length of time spent participants spent on each level of difficulty. While participants spent increasing lengths of time as level difficulty increased, there were individual variation within levels, suggesting that the system was responding to user cognitive workload for each participant individually. 65

5.1 The musical harmonies added by BrAAHMS. a) Blue notes denote the keys the user is pressing. Red notes denote the additions made by the system. The L and R depict the fingers of the left and right hand respectively. The * indicates middle C. b) Music notation showing a sample user input (blue) with the additional pitches provided by the system (red). 75

5.2 Sound engine behind BrAAHMS. Musical additions are implemented in Pure Data through a piano soundfont on FluidSynth. Musical additions were only added or removed when a marker was received from the fNIRS system. 76

5.3 Notated music provided as guidance only to participants during musical improvisation. 77

5.4 Left: Mean change in oxy-Hb and standard error across all trials and participants. Although each participant was modeled individually, the fNIRS signal exhibited a general trend with higher levels of oxy-Hb corresponding with hard pieces, consistent with the findings of Experiment 1 (Figure 3.10). Right: mean change in oxy-Hb was significantly higher in participants when they played an hard piece on the piano versus an easy piece consistent with the findings of Experiment 1 ($p < .001$, also see Figure 3.10). 81

5.5 Favorite conditions of 20 participants 82

5.6 Rank-Sums of the conditions based on a median split (6.5 years) on years of experience playing the piano. Rankings of participants with lower experience (< 6.5 years, $n = 10$) (left) and and higher experience (> 6.5 years, $n = 10$). Lower bars indicate a more preferred condition, and bars with the same letter are not significantly different. 86

6.1 Highly attractive agents (left) courtesy of Dr. Stuart Gibson and Dr. Chris Solomon, University of Kent. Neutrally attractive agents (right) courtesy of Dr. Martin Gruendl, Beautycheck. 91

6.2 Mean and standard error of participants' ratings of the highly and neutrally attractive agents' *left*: physical attractiveness, *center:* vocal attractiveness, and *right:* vocal human-likeness. 93

6.3 Top: Example of question without a suggestion. The time remaining can be seen on the bottom-left. Bottom: Example of a question with an agent offering a suggestion. Participants would hear the agent offering a suggestion as well. If they wished, the participant could hear the suggestion by pressing on the *'Tell me the answer'* button. 95

6.4 Examples of feedback for an incorrect (left) and correct (right) answer. Participants gained a point for every correct answer, and lost a point for every incorrect answer and every time they timed out in order to discourage guesswork. 96

6.5 Mean and standard error of users' self-reported trust in agents, user perception of agent accuracy, number of correct suggestions accepted, and time spent on questions with suggestions. 104

7.1 The valence-arousal model of emotion. 115

A.1 Easy Piece 1 145
A.2 Easy Piece 2 146
A.3 Easy Piece 3 146
A.4 Easy Piece 4 146
A.5 Easy Piece 5 147
A.6 Easy Piece 6 147
A.7 Easy Piece 7 147
A.8 Easy Piece 8 148
A.9 Easy Piece 9 148
A.10 Easy Piece 10 148
A.11 Easy Piece 11 149
A.12 Easy Piece 12 149
A.13 Easy Piece 13 149
A.14 Easy Piece 14 150
A.15 Easy Piece 15 150
A.16 Hard Piece 1 150
A.17 Hard Piece 2 151
A.18 Hard Piece 3 151
A.19 Hard Piece 4 152
A.20 Hard Piece 5 152
A.21 Hard Piece 6 153
A.22 Hard Piece 7 153
A.23 Hard Piece 8 154
A.24 Hard Piece 9 154
A.25 Hard Piece 10 154

A.26 Hard Piece 11 155
A.27 Hard Piece 12 155
A.28 Hard Piece 13 156
A.29 Hard Piece 14 156
A.30 Hard Piece 15 157

B.1 Consent Form for Expts 1, 2, and 3, Page 1 159
B.2 Consent Form for Expts 1, 2, and 3, Page 2 160
B.3 Questionnaire for Expts 2 and 3, Page 1 161
B.4 Questionnaire for Expts 2 and 3, Page 2 162
B.5 Questionnaire for Expts 2 and 3, Page 3 163
B.6 Questionnaire for Expt 2, Page 4 164
B.7 Questionnaire for Expt 3, Page 4 165
B.8 Questionnaire for Expt 4 166
B.9 Questionnaire for Expt 4 167
B.10 Questionnaire for Expt 4 168

List of Tables

4.1 Results from Wilcoxon Signed-rank test. Significant results are highlighted in bold. Findings indicate a pattern for increased accuracy and speed in BACh over the normal condition. 60

6.1 Results: 2x2 RM-ANOVAS on agent reliability and attractiveness. . 102

Chapter 1

Introduction

1.1 Motivation

While there have been a great deal of technological advances in computer science in the last decades, the methods and interfaces used to interact with computers still remain the QWERTY keyboard and mouse. The touch-screen has not changed this, whereby using a finger can be equated directly to using a mouse. These relatively antiquated methods are still the main instruments of human-computer interaction. Yet, the QWERTY keyboard was patented in 1878 [193] and the mouse was first prototyped as the trackball or roller ball in 1941 and patented in 1947 [173].

While both the human and computer are complex machines capable of sophisticated functions, this limitation in interaction has created an extremely limited bandwidth for information to pass between the two. The human can type out commands on a keyboard and click on a mouse (and occasionally try to give voice commands which may or may not be misunderstood). In return the computer can display information on pixels on a 2-dimensional screen (and occasionally translate the information vocally).

Due to the acceptance of the status quo, we do not treat our computers as our peers, we barely trust them to be our assistants. Yet, computers hold the potential to respond intelligently to us if only they could have access to more information about us.

In this thesis, I will address the topic of the next generation of intelligent brain-computer interfaces that provide the computer with more information about the human (with no additional effort on the part of the user) so that the computer can respond more intelligently in return.

1.1.1 Why Brain-Computer Interfaces?

Brain-computer Interfaces (BCIs) have traditionally been developed as a replacement for the mouse and keyboard for severely physically disabled people (e.g., [165, 98, 109]. Such BCIs use explicit brain signals to carry out commands such as selecting a target out of a grid of options or moving a cursor in different directions. This topic will be discussed further in Chapter 2. Such traditional BCIs are too slow and too inaccurate for the general population, however, and as such, are not a viable method of human-computer interaction for most people [147].

Brain-computer interfaces are ideally poised to act as *augmentors* of the traditional limited channel of communication between the human and computer, rather than as an *alternative*. This is due to a combination of factors.

Firstly, BCIs provide a *direct* measurement into a user's state such as cognitive workload. Other physiological measures such as electrodermal activity (EDA), heart rate variability (HRV) or respiration rate are indirect measures of user state as they measure a secondary response from the brain. For example, HRV has been linked to cognitive performance and sustained attention.

Secondly, BCIs provide an *objective* measurement into user state as opposed to questionnaires or behavioural analysis which requires subjective rating. The former is also prone to bias.

Thirdly, BCIs provide a continuous flow of information in real-time in contrast to questionnaires which are measurements of discrete points in time. This is especially relevant for the purposes of adaptive, intelligent systems that can respond to user state in real-time.

In this way, if BCIs can be used as *an additional channel of communication in addition to the bread-and-butter keyboard and mouse techniques*, they could augment

the limited bandwidth of human-computer interaction.

In recent years, a new generation of BCIs has started to emerge that is designed for the general population [12, 199, 159, 160, 168, 3, 4, 77, 119, 115]. Such BCIs use *passive* brain signals, i.e. the user carries out their task as usual, and the interface detects their cognitive state implicitly in the background. Some of these BCIs also respond in real-time with an adaption. Such adaptations have generally been used for the purposes of *automation of specific tasks* such as a virtual robot in a search task [199], the bubble cursor target size [4] or the number of unmanned aerial vehicles in a ground control simulation [82, 3].

In this thesis, I move away from adaptive BCIs that automate tasks, and towards fields that require rather different responses. I build and expand on the next generation of BCIs with two novel real-time, adaptive, intelligent systems where users' cognitive workload is the *determining factor* in ***learning*** and ***creativity***. These are both previously unexplored fields in BCIs and present their own individual challenges.

There has yet to be a real-time, adaptive intelligent learning system that responds to learners' physiological measures of cognitive workload. Cognitive load theory has highlighted the lack of an objective measure in cognitive workload [34] which is a crucial part of cognitive load theory and instructional design material. I address this challenge in Chapter 4 with the system **Brain Automated Chorales (BACh)**.

In a parallel vein, creativity has been flagged by the NSF Workshop on Creativity Support Tools as a "grand challenge" for HCI researchers [191, 192]. Mayer [122] highlighted how brain data could provide information not available to other methodologies in the field of creativity. Yet, as pointed out by Mayer [122] and echoed by Shneiderman et al. [191, 192], it is not yet clear *how* an intelligent, adaptive system that measures cognitive workload could provide creative support to a user. That is precisely the problem I address in Chapter 5 with the system **Brain Automated Adaptive Harmonies in a Musical System (BrAAHMS)**.

Learning and creativity are also both vast fields that I narrowed down through

the medium of playing a musical keyboard.

The topic of music was ideally situated for both learning and creativity. I used a musical learning task on the piano keyboard because it lends itself well to task segmentation with increasing levels of difficulty and to high element interactivity due to concurrency. A musical task is also easy to evaluate in terms of accuracy and speed when compared with a control condition.

In terms of creativity, a musical improvisation task lends itself well to adaptations that are harmonious with what the user is currently playing and demands a high level of real-time precision.

Before building the two novel BCIs for the purposes of learning and creativity, I first demonstrated that it was possible to measure user cognitive workload while users played the piano.

After having evaluated both systems and demonstrated that they did indeed improve user learning and creativity, I then proceeded to investigate factors that would enable users to trust and use such intelligent systems. It is not unlikely that before long intelligent agents will become common in the fields of learning/education and creativity. Pedagogical virtual agents and robotic agents are already being investigated as educational tutors, e.g., [44, 137, 182, 178]. Virtual agents are also becoming increasingly used in situations requiring creative abilities, e.g., [7, 13, 126, 33]. It is therefore fitting that we investigate what factors will encourage users to interact with such agents and start treating them as peers, to prepare the way for when learning/education and creativity will be strongly influenced, if not dominated by, intelligent, embodied agents. After all, what is the point of creating intelligent systems that could respond to humans as peers, if humans will not treat them as such?

I will discuss the thesis contributions in more detail in the next section.

1.2 Thesis Contributions

As discussed above, this new generation of passive BCIs have previously explored adaptations involving automation of tasks. In this thesis, I explore the use of user cognitive workload as the *determining factor* in increasing user learning and creativity. I then explore factors to increase trust in such intelligent systems. As such, I carried out four experiments with the following contributions:

- *Experiment 1*: Discovery and demonstration that high and low cognitive workload can be measured using fNIRS when users play the piano.
- *Experiment 2*: Build an adaptive, interactive musical learning system, BACh, that guides learners' progress by measuring when they can cognitively handle more information and providing them with the next stage of learning at the right moment.
- *Experiment 2*: Demonstrate that users could play the musical pieces faster and with more accuracy using BACh than they normally would in a control condition.
- *Experiment 2*: Demonstrate through questionnaire and interview data that participants felt they learnt the pieces better with BACh and they liked timings of the level changes.
- *Experiment 2*: Implement difficulty level specific parameters to real-time data classification.
- *Experiments 2 and 3*: Implement individual confidence threshold automation for additional data classification accuracy of brain signal rather than fixed percentage thresholds.
- *Experiment 3*: Iterative design and demonstration of a novel musical instrument system, BrAAHMS, that uses cognitive workload in real-time to add and remove musical harmonies during musical improvisation on the piano.

- *Experiment 3*: Experimental evidence from interviews and subjective rankings show that users prefer BrAAHMS to other conditions because they felt more creative and that the system was responsive to them.

- *Experiment 4*: Build a trust evaluation system containing virtual agents with differing levels of reliability and attractiveness in a QA search task.

- *Experiment 4*: Demonstrate through behavioural and questionnaire data that reliability alone is not enough for users to trust a virtual agent, and that attractiveness is perhaps more important.

1.3 Thesis Overview

In light of the above contributions, this thesis is structured accordingly:

1. In Chapter 2, I carry out an overview of previous work in the field of brain-computer interfaces, learning, creativity, and trust.

2. In Chapter 3, I demonstrate through a **feasibility study** that it is possible to accurately measure cognitive workload using a fNIRS-based BCI while users play a musical keyboard. This work sets the foundations for applying fNIRS to the fields of learning and creativity using the medium of playing the piano.

3. In Chapter 4, I apply fNIRS to a real-time musical learning task using **BACh (Brain Automated Chorales)**. BACh guides learners progress by measuring when they can cognitive handle more information and providing them with the next stage of learning at the right moment. I show that users learned with increased speed and accuracy using BACh over a control condition.

4. In Chapter 5, I then apply fNIRS to aid users in the real-time creative task of musical improvisation with **BrAAHMS (Brain Automated Adaptive Harmonies in a Musical System)**. I demonstrate that users felt more creative using BrAAHMS than with control conditions.

5. In Chapter 6, I build a **trust evaluation system** that explores factors important to users to trust intelligent systems. I demonstrate that attractiveness is just as important, if not more so, than virtual agent reliability.

Chapter 2

Related Work

2.1 Brain-Computer Interfaces

2.1.1 Traditional Approaches

A brain computer interface (BCI) was defined in the first international meeting on BCI technology as "... *a communication or control system in which the user's messages or commands do not depend on the brain's normal output pathways of peripheral nerves and muscles.*" Wolpaw et al. [222].

This can allow people with severe motor impairments such as amyotrophic lateral sclerosis (ALS), brainstem stroke, spinal cord injury, muscular dystrophies and severe cerebral palsy to be able to have an alternative way of interacting with a computer in real time. In some cases people can lose all voluntary muscle control including eye movement such as totally locked-in patients which excludes eye gaze technologies as a form of communication. A non-muscular method of control from the brain directly to the computer can be an option for restoring communication with the external world.

2.1.1.1 BCI Techniques

Electroencephalography (EEG) BCIs are commonly used due to the fact that they are generally considered the cheapest and least complicated type of BCI (e.g., [19,

165, 98, 109, 224, 223]. They have high temporal resolution but poorer spatial resolution due to the fact that measurements are taken from the scalp. Methods other than EEG used in BCIs include magnetoencephalography (MEG), positron emission tomography (PET), functional magnetic resonance imaging (fMRI) and optical imaging. Examples of work done in these areas include Weiskopf et al. [218] who trained subjects to control the blood oxygen level-dependent (BOLD) response of two areas of the brain (the supplementary motor area (SMA) and the parahippocampal area (PPA)) using an fMRI-BCI with feedback within 1.3 seconds. In Yoo et al. [228] subjects were able to navigate a 2D maze using a fMRI-BCI method. An example of a MEG-BCI is given in Mellinger et al. [131] where subjects were trained to communicate binary decisions by the imaginary movement of limbs.

An emerging BCI technique used in HCI is functional near infrared spectroscopy (fNIRS), an optical brain sensing technique that measures level of activation in the brain due to the hemodynamic response. FNIRS has been used to measure levels of cognitive workload in the prefrontal cortex (e.g., [197, 12, 199, 160, 168, 3, 4, 119, 77]).

BCIs can be broadly classified into two categories depending on the type of input signals: operant conditioning based BCIs and evoked potential BCIs.

2.1.1.2 BCI methods requiring training

BCIs based on operant conditioning depend on direct user control of certain frequency bands in particular cortical areas. This approach requires training which can last up to several months for a patient with ALS [19] and it has been found that 1 in 5 subjects cannot be trained [224]. An advantage though is that the user can control the device faster over time as they learn to control specific EEG signals. One of the methods used in endogenous BCI is the alteration of the amplitudes of mu (μ) and beta (β) rhythms in the sensorimotor cortex due to the imaginary movements of the extremities [165, 98, 109]. Kostov and Polak [98] trained subjects to move a 2D cursor on a monitor using UP and DOWN-thoughts. Leeb et al. [109] showed that it was possible for subjects to move through a fully immersive virtual environment

using imaginary foot movement. Another approach used in endogenous BCI is to use a wider distribution of EEG signals that are not movement related. It has been shown that subjects can change their μ rhythm amplitudes to control a 2D cursor on a monitor [224, 223].

When someone is not engaged in movement or in the sensory processing of movement, they typically exhibit brain wave oscillations in the α-band frequency of 8-1 2Hz in the sensorimotor areas of the brain [164]. These frequencies, which can be detected by EEG, are called μ rhythms. These μ rhythms are thought to be produced by the hand and foot areas of the sensory and motor cortices.

Motor imagery can cause changes in the sensorimotor rhythms without actual movement taking place, thus, it can serve as a tool used in BCIs. There have been many studies carried out on motor imagery-based BCIs such as Kostov and Polak [98] who created a parallel man-machine training approach where users could be trained to control a cursor in 1 (up-down) and 2-dimensions(up-down-left-right). Their two participants were able to hit 1D targets with 100% accuracy and 2D targets with 63% and 73% accuracies. Recently, McFarland created a 3D cursor, which with extensive training can be controlled by combinations of different motor imagery commands [127].

Motor imagery-based BCIs have also been used to control wheelchairs. For example, Galan et al. [61] used left hand imagination movement, rest and word association to move physical and virtual wheelchairs. However only the experiments with the virtual wheelchair were reported where subjects were able to reach 100% and 80% of their goals. Robots are also able to be controlled by motor imagery based BCIs. Millan et al. [134] were able to control a mobile robot to navigate through several rooms. Mental control was compared to manual control with a performance ratio of 0.74.

One of the key aspects of motor imagery-based BCIs is that there must be a stable temporal response in the μ and β-bands with little intrasubject variability. However, differences have been found in the dynamics of oscillatory activity [163]. It is thought that these differences may be due to the great range of motor imagery. For

example, a user may imagine themselves performing the action from the first person point of view, which would involve kinesthetic feelings. Or a user may imagine themselves or another person performing the action from a third person point of view, which is more visually based. Neuper et al. [147] investigated the influence of different kinds of imagery: they looked at kinesthetic motor imagery from the first person (MIK) and visual motor imagery from the third person (MIV) (as described above). They also had two control conditions: actual movements (ME) and just visual observation of hand movements with no imagery (OOM). Results showed that the highest classification accuracies (around 80%) were for the real conditions (ME and OOM). The MIK delivered an average classification accuracy of 66% and MIV gave 56%. Interestingly, for both ME and MIK, electrodes close to position C3 gave the best input features, whereas for MIV, the parieto-occipital cortical areas were the most active. This work suggests that specific instructions to users on how to imagine actions can enhance activation of the sensorimotor areas, thereby improving motor imagery-based BCIs.

2.1.1.3 Evoked Potential BCIs

The second category of BCIs are those that are based on an evoked potential which is an increase in amplitude in the EEG signal due to a stimulus. No training is required for this type of BCI. To increase the signal-to-noise ratio (SNR) signal averaging is usually required. This category contains different types of input signals including steady-state visual evoked potentials (SSVEP) and P300 evoked potentials.

In a SSVEP system there are flickering light stimuli. The stimulation frequency of one of the stimuli can be found by the user looking at that stimulus which elicits SSVEPs of the corresponding frequency in the visual cortex. By using two flickering sources McMillan et al. [129] showed that the roll position of a flight simulator could be controlled. Friman et al. [60] used SSVEPs to create a BCI to spell words with an average information transfer rate of 27 bits/minute and a 97.5% accuracy rate. SSVEPs are slightly different to the P300 evoked potential where the user can focus on one particular auditory or visual stimulus out of many and for

example may count the occurrences of that stimulus. This causes a P300 waveform in the EEG signals around 300 ms after the event.

The P300 waveform is an event related potential (ERP) which is an increase in voltage about 10 μV peaking around 300ms after the stimulus. It is triggered by an auditory, visual or somatosensory stimulus which is infrequent or particularly significant amongst other more routine stimuli, otherwise called the "oddball" paradigm [167]. The P300 is made up of 3 components: P3a (peaking around 250ms), P3b (peaking around 350ms) and the slow wave [200]. The P3a is more prevalent in the frontal area than the P3b whilst the slow wave is clearer in the parietal areas. The P3b and slow wave increase in amplitude when the subject pays attention to the stimuli but the P3a does not [200].

As the P300 wave is maximally recorded from the midline centroparietal area, at least the scalp locations Fz, Cz and Pz in the international 10-20 system are needed and preferably more [167]. However it is to be noted that after a review of many studies Johnson [85] reports that the scalp distribution of the P300 varies greatly. This suggests that either there is one basic P300 wave which is affected by different experimental conditions or that there may be several different generators of the P300 [167].

As the difficulty of the task to identify the stimulus increases, the amplitude of the P300 decreases and the latency increases. This has been explained by a drop in confidence of the user in identifying the correct stimulus [200]). In general as the users age increases there is a drop in the P300 amplitude and an increase in latency. It is not clear where the sources of the P300 waves are and they are likely to be multiple which vary according to the task and the individual. It is also not clear as to what the purpose of the P300 wave is likely to be. However the P300 wavefrom has been used successfully in a number of BCI studies.

The first P300 BCI was carried out by Farwell and Donchin [52] where subjects were presented with 6 x 6 matrix of letters and numbers. The rows and columns were flashed one at a time and subjects were asked to keep a running mental count of the flashes of the letter they wanted to spell during each set. Each time the

desired letter flashed and the subject counted a P300 waveform was triggered so that the letter and P300 could be correlated in time. Results showed that this communication channel could operate at 12 bits/minute, or 2.3 characters/minute with a character being chosen with 95% accuracy within 26 seconds. This speed was improved upon with Donchin et al. [43] with 7.8 characters/minute with 80% accuracy. The 6x6 grid was enhanced to a 8x8 grid with a multiple flash approach with half of the stimuli being flashed simultaneously producing stronger ERPs with faster speeds [6]. This may well be a more efficient approach to the traditional P300 speller. More recently, Yuksel et al. [232] got higher classification accuracies by extending the traditional matrix onto a touch-table with physical objects.

2.1.2 Next Generation of Implicit Brain-Computer Interfaces

In addition to the two broad categories of BCI described above, there is an emerging new generation of implicit BCIs. This new generation of BCIs could be categorized as "passive" under Zander et al.'s [235, 234] categorization of BCIs:

1. *Active BCIs* are driven by explicit brain signal input under conscious control, such as the motor imagery BCIs described in section 2.1.1.2 (e.g., [98, 108]).

2. *Reactive BCIs* are driven by brain activity that react to external stimulation unconsciously, such as the P300-based or SSVEP-based BCIs in section 2.1.1.3.

3. *Passive BCIs* are driven by implicit brain signal input unconsciously. They have been defined as operating "without the purpose of voluntary control, for enriching a humanmachine interaction with implicit information on the actual user state" [234].

The third category, passive BCIs, have also been termed *implicit* BCIs, i.e. signals that users naturally give off when interacting with interfaces [30, 208].

While active and reactive BCIs can be invaluable to people with motor disabilities, to the general population they are slower and less accurate than their standard mouse and keyboard. Recently, however, fNIRS and other brain-sensing

technologies have been used for passive or implicit brain computer interfaces, where the user performs a task normally, but the user's brain signals act as an *additional* communication channel that are used to assess cognitive state and adapt the interface accordingly.

One of the most promising signals to control real-time BCIs is cognitive workload, which has been used since the first fully adaptive loops [172, 221], and has been shown to successfully adapt an interface using fNIRS [3, 67]. In addition, fNIRS has also been used to control passive BCIs using preference [159] and multitasking signals [4, 199].

Multi-tasking has also been demonstrated by fMRI [96], EEG and fNIRS [198, 4]. Preference has also been demonstrated with fMRI [158, 125, 35] and fNIRS [159]. Such findings have propelled BCI research forwards into the development of real-time BCIs based on user cognitive state. For example, Solovey et al. [199] automated virtual robots in real-time based on multitasking brain activity patterns. Afergan et al. used the same measurements to adapt bubble cursor target size[4]. Cognitive workload has been used in unmanned aerial vehicles ground controller simulations [82, 3]. Preference has been used to select movies for users [159]. This new generation of of BCIs offer a new step forward from the more traditional BCIs that were designed for severely physically disabled users. In the next sections, I will discuss how these new types of BCIs measure user cognitive workload in the prefrontal cortex.

2.1.3 Cognitive Workload and the Prefrontal Cortex

The working memory model proposed by Repovš and Baddeley [177] is made up of multiple components: a central executive, a visuospatial sketchpad, a phonological loop, and an episodic buffer. The phonological loop contains verbal information in the form of a short-term acoustic store and an articulatory rehearsal process. The visuospatial sketchpad is comprised of visual and spatial subsystems with independent storage, maintenance and manipulation processes. The central executive seems to have several different functions which are involved in the manipulation of infor-

mation within the memory stores. The latest addition to the model, the episodic buffer, holds and integrates information from other working memory components and from long-term memory into scenes or episodes. For an in-depth description and discussion of the multi-component model of working memory please see Repovš and Baddeley [177].

Functional magnetic resonance imaging (FMRI) and positron emission tomography (PET) studies have established that cognitive workload can be measured in the prefrontal cortex e.g. [36, 121]. The prefrontal cortex is important for higher-order cognitive functions including working memory [174]. The dorsolateral prefrontal cortex has been associated with working memory and with high-level organization of the contents of working memory. Bor et al. (2003) presented subjects with a spatial working memory task with structured and unstructured sequences.

The ventrolateral prefrontal cortex is also associated with working memory and is thought to be linked to the retrieval of a few pieces of information. For example, Jonides et al. [87] gave participants a memory task by presenting three stimuli concurrently for 200ms. They then asked participants whether a probe circle displayed 3 seconds later occupied one of the same locations. Using positron emission tomography (PET), Jonides et al. [87] found increased activation in the mid-ventrolateral prefrontal cortex in contrast to a simple perception condition. These findings suggest that the intention to remember may be the cause of the activation in the ventrolateral cortex.

One area of the prefrontal cortex that is not so widely understood but is also associated with working memory is the anterior prefrontal cortex (see [174] for a review). This constitutes Brodmann area (BA) 10 and is very well developed in humans in comparison to other primates. Some of the neuroimaging evidence points towards functionalities such as memory retrieval, retrieval verification and source memory [181]; prospective memory such as "at time x, do y" [21]; cognitive branching or the ability to hold goals while carrying out secondary goals [96]; and relational knowledge [99].

A theory put forward by Ramnani and Owen [174] is that the anterior pre-

frontal cortex is activated when a problem requires more than one cognitive process. Hence, the cognitive operations need to be coordinated in the pursuit of a more general goal. Their hypothesis is broadly consistent with much of the previous functional neuroimaging literature described above, and with the theory of the central executive of Repovš and Baddeley's (2006) model [174].

The prefrontal cortex has also been imaged by fNIRS, which is the BCI technique used in this thesis, and is discussed in greater detail below.

2.1.4 functional Near Infrared Spectroscopy (fNIRS)

Functional near infrared spectroscopy (fNIRS) is an optical brain sensing technique that measures levels of activation in the brain due to the hemodynamic response. It is relatively resistant to movement artifacts [197], and observes similar physiological parameters to functional magnetic resonance imaging (fMRI) [24].

fNIRS uses near-infrared light to measure concentration and oxygenation of the blood in the tissue at depths of 1-3cm [213]. Light is sent into the forehead in the near infrared range (650-900 nm), where it is diffusely reflected by the scalp, skull, and brain cortex. At this wavelength, oxygenated and deoxygenated hemoglobin are the primary absorbers of light. A very small percentage of the light sent into the head returns from the cortex to the detector on the fNIRS probe. By measuring the light returned to the detector, researchers are able to calculate the amount of oxygen in the blood, as well as the amount of blood in the tissue.

Biologically, when a region of the brain is active, there is an increase of blood flow to that region [36]. This increase of blood flow is typically coupled with increased levels of oxygenated hemoglobin [55]. Thus, fNIRS can be used to measure activity in localized areas of the brain.

To make this calculation, raw data can be transformed into hemoglobin concentrations using the modified Beer-Lambert Law [5]:

$$\Delta A = \varepsilon \times \Delta c \times d \times B \tag{2.1}$$

where ΔA is the change in attenuation of light, ε is the molar absorption coefficient of the absorbing molecules, Δc is the change in the concentration of the absorbing molecules, d is the optical pathlength (i.e., the distance the light travels), and B is the differential pathlength factor. The attenuation of light is measured by how much light is absorbed by oxygenated and deoxygenated hemoglobin (which are the main absorbers of near infrared light at these wavelengths). As the attenuation of light is related to the levels of hemoglobin, given ΔA, we can derive the changes in the levels of oxygenated and deoxygenated hemoglobin [24].

By placing the fNIRS probes on the forehead, brain activity in the anterior prefrontal cortex can be measured. This placement offers the advantage of fNIRS measurements without the interference of hair follicles, which can absorb light and disrupt the signal.

Functional magnetic resonance imaging (FMRI) and positron emission tomography (PET) studies have established that cognitive workload can be measured in the prefrontal cortex [36, 121]. More recently, brain imaging studies have examined changes in brain function associated with improvements in performance of tasks [63, 171, 15, 92].

We can therefore use fNIRS to measure levels of cognitive activation in the anterior prefrontal cortex (PFC) by placing sensors on the forehead. The PFC is the seat of higher cognitive functioning such as complex problem solving and multitasking [96].

The fNIRS signal has been found to be resilient to respiration, heartbeat, eye movement, minor head motion, and mouse and keyboard clicks [197]. It is generally more tolerant of motion than EEG and has a higher spatial resolution. However it does have a slower temporal resolution than EEG with a delay of 5- 7 seconds due to the hemodynamic response of blood flow to the brain. Due to its general ease in setting up with users and its relative tolerance of minor motion, fNIRS is an increasingly popular method of brain sensing in the HCI community [197, 12, 199, 160, 168, 3, 4, 119, 77].

2.2 Learning

As discussed in the Introduction, the two main fields of application in this thesis are *learning* and *creativity*. I also then go on to explore user trust in intelligent systems. As such, I will cover the related work for each system.

As background literature for Experiment 2, Chapter 4, I shall now discuss learning in the context of cognitive workload and adaptive, intelligent systems.

2.2.1 Computer-Based Education

Computer-based education (CBE) is a generic term that is frequently used to encompass a broad spectrum of interactive, adaptive computer applications used in teaching [84]. I will use the term CBE as the broadest umbrella term in this thesis, covering all methods of computer aided pedagogy. A closely related term is computer assisted instruction or computer aided instruction (CAI). Other terms include computer-based instruction, computer-based learning amongst others and there have not been clear cut definitions or agreeance across all terminologies. CAI, and sometimes CBE, is considered to be distinguishable in further sub-areas such as: (a) *Drill-and-practice* applications, (b) *Tutorial* mode, (c) *Simulation* mode, and (d) *Instructional game* [84].

A more specific subset of computer applications used in teaching is covered by the term intelligent tutoring systems (ITS). These systems are delineated by having knowledge of: (a) *the domain* by having the expert model for the subject, (b)*the student* and following their progress, (c) *tutor model* by making teaching choices based on the student and the material or (d) *user interface model* to tie all models together [151, 149].

CBE has shown that students tend to learn more, take around 30% less time and have increased enjoyment of their subject [100]. This results in students feeling more successful, leading to greater motivation to learn and increased self-confidence [17].

In this thesis, I developed BACh, based on measuring and adapting to learn-

ers' cognitive workload using brain sensing, under the broad category of CBE. I do not profess to our system meeting all the requirements of the definition of an ITS, instead, however, BACh certainly brings benefits to the field of ITS by providing knowledge about the learner that can sway an ITS's teaching choices.

Since the measurement of cognitive workload is an integral part of BACh, I now discuss the foundations of cognitive load theory and how it relates to a real-time, BCI-based CBE.

2.2.2 Finite Cognitive Workload and Learning

The fundamental idea behind both Baddeley's working memory model [177] and Cognitive Load Theory (CLT) [204] is that cognitive workload is limited in its capacity to handle information. CLT applies this finite, limited cognitive capacity of working memory to instructional design in order to minimize any unnecessary burdens on working memory and maximize the formation of automated schemas in long-term memory (i.e. learning). CLT uses schema theory [28] whereby a schema is a cognitive construct that allows us to treat multiple elements of information as a single element. A classic example is our classification of all trees as the schema 'tree' even though each tree is unique from every other tree [204]. The functions of schemata are: 1) to allow the means of storing huge amounts of information in long-term memory, and 2) to vastly reduce the burden on working memory during tasks. For example, as children learning to read, we had to build schemata of letters, then words, then combinations of words and phrases. We gradually built increasingly complex schemata from lower-level schemata. This allows us to devote our working memory to the more important task of considering meanings and ideas when we read today which would otherwise be impossible if we were having to use our finite cognitive capacity to decipher each letter from scratch.

The degree of element interactivity is seen to be the most important characteristic that determines the complexity of learning material. CLT says that an expert differs from a novice because they have previously acquired schemata that have built complex elements into fewer elements. In a task, the novice will have

to use more of their limited working memory access and process the individual elements than the expert, reflecting in the difference in skill. In the example of a pianist, a skilled musician can cognitively group together notes to form chords in their schemata, whereas a beginner has to process each note individually, suggesting that the beginner will have a higher cognitive workload while learning the same piece as the skilled pianist.

Part-task approaches have been investigated to generate instructional design that does not overload the learner. However, artificially reducing cognitive workload by making the learning material easier also reduces understanding. Therefore it is important to know when to integrate all the parts that have been learned separately. Meta-analyses have found part-task training is only helpful with sequential material, such as playing sequential notes on a piano (e.g. [11]), rather than concurrent tasks, such as playing two hands on a piano at the same time [219, 220]. The missing element in part-task training strategies for this type of learning task is thought to be the opportunity to practice time-sharing skills [32, 220, 219] to allow for the multitask fluency. However, it has been found that increasing-difficulty tasks *were* helpful if the tasks adapted to the learners individually and not with fixed scheduling [219].

These findings highlight the importance of adaptive learning tasks, especially when using increasing-difficulty strategies. In order for increasing-difficulty training to be adaptive, there has to be some measurement of when the user is ready to move on to the next level of difficulty.

2.2.3 Measurement of Cognitive Workload in Learning Studies

The method of cognitive workload measurement in learning studies have come under criticism and scrutiny in recent years [20, 34, 144, 90]. Moreno [144] has stated that "insufficient attention has been given to the rigorous development of [cognitive load] measures". Brunken et al. (2003) state that "despite 3 decades of [CLT] research, its key concept - cognitive load - has largely eluded our attempts of direct measurement". In fact, in a paper titled "Contemporary cognitive load theory

research: the good, the bad and the ugly", the "ugly" refers to cognitive workload measurement.

Cognitive workload in CLT studies have typically been measured in one of three ways: self-reporting, secondary-task methodology, and physiological measures.

2.2.3.1 Self-reporting questionnaire

The most frequently used measure is the self-reporting questionnaire created by Paas et al. (1992) [155] which is a one-item scale to measure "perceived amount of mental effort" [155]. De Jong [34] and Moreno [144] raise several issues with the scale's sensitivity, reliability and validity in measuring cognitive workload. Firstly, it is not a direct measurement of cognitive workload, it is measured, instead, by interpretation of post-tests [34]. The need for direct measurement of cognitive workload has been highlighted by Mayer (2005) [123]. In fact, more objective measures of cognitive workload using brain sensing have shown discrepancies between self-reported levels and brain activation levels [77].

Another important factor is the timing of the questionnaires, which varies across studies [34]. Many studies only present the questionnaire after learning has taken place, so it is not clear if learners provide an average estimate for the whole task, and it leaves reports open to memory and consciousness effects. However, cognitive workload is a dynamic measure that fluctuates during a learning task [227].

2.2.3.2 Secondary tasks

A second method used to measure cognitive workload in CLT studies has been the use of secondary-task or dual-task measures (e.g. [203, 20]). A secondary task is carried out concurrently with the primary learning task which is measured in performance. The idea behind this methodology being that the secondary task is supposed to measure the cognitive workload induced by the primary task. As the primary task requires more cognitive workload to accomplish, performance in the secondary task is supposed to wane. An important disadvantage to this methodology is that

the secondary task can interfere considerably with the primary task, especially if the primary task is complex and requires much of the learner's limited cognitive capacity [153, 20]. Furthermore, Brunken et al. [20] points out that the sensory modality of the dual tasks will affect the outcome. According to Baddeley's working memory model [177] different components of working memory use different sensory inputs. However, the different components are not completely independent and still affect each other.

2.2.3.3 Physiological Measures

Another set of methods that has been used to measure cognitive workload in learning studies is physiological measures. However, most physiological measures of cognitive workload have only been analyzed offline. This misses the advantage that physiological measures provide which is the continuous availability of data which is especially important for adaptive tasks that can react to learners' needs in real-time.

Heart rate variability (HRV) and electrodermal activity (EDA) (previously called galvanic skin response (GSR)) have been shown to be affected by cognitive workload [57, 101, 186]. However, changes in physiological measures such as HRV and EDA can also be mapped to other phenomena [51] such as changes in emotional state. Paas and van Merrienboer [154] measured heart rate variability (HRV) as well as the self-reporting scale while users carried out different conditions in a geometrical problem solving task. They did not find any differences in HRV between the conditions and thus concluded that this measure was "invalid and insensitive" [153]. However, more recent work gives support to HRV as a measure of cognitive workload. For example, differences in heart rate were found to reflect levels of cognitive workload during simulated combat mission flights [101] and a combination of heart rate and systolic blood pressure was reported as an objective measure to separate cognitive workload and cognitive stress in three cognitive tasks [57]. HRV was found to be significantly different during high mental workload induced by a complex flight in novices than during recovery over a 24 hour period [185].

Task-evoked pupillary responses (TEPRs) have also been associated with

changes in cognitive workload [212, 81]. However, pupil size is sensitive to a number of factors that are not relative to cognitive workload such as the ambient light in the environment [187]. TEPRs have not been used in many learning studies [34] and therefore its sensitivity to learning tasks has not been fully established. This is especially important as pupil size is sensitive to a number of factors that are unrelated to cognitive workload such as the ambient light in the environment. Studies using TEPRs as a measurement of cognitive workload have to control for constant lighting, avoidance of eye movements, use of non-visual stimuli, and use of shorter tasks [187], all of which lead to unrealistic conditions for learning. For example, Iqbal et al. [81] tested TEPRs with cognitive workload in a HCI study consisting of four different tasks. Interestingly, while task completion times and user ratings indicated differences in easy and hard condition in all four tasks, there were no differences in TEPRS in two of the tasks: reading comprehension and file management. The authors then decomposed these tasks into parts and found some differences, which suggests TEPRs may only be sensitive to cognitive workload during certain periods of a task. Schultheis and Jameson [187] also tested continuous reading and TEPRs along with reading speed, subjective load and P300 amplitude. They also found that the differences in cognitive workload were not observable in pupil diameter whereas the other measures did pick up differences between easy and hard reading texts. It is important to report such negative results because it suggests that the effectiveness of TEPRs at measuring cognitive workload can be strongly dependent on the type of learning task and the situation in which it is used.

2.2.3.4 Cognitive Workload Brain Imaging

Brain sensing or neuro-imaging techniques have been discussed in many studies as a promising field for cognitive workload measures during learning [153, 20, 187, 34]. Brain sensing provides a more direct measurement of the physiological changes occurring in the brain during higher cognitive workload and can be measured continuously which can be used as a real-time determinant in adaptive learning studies. However there have been very few studies to take advantage of this measurement

tool.

EEG has been shown to measure cognitive workload via changes in theta and alpha range signals (e.g. [75, 66, 65]). Gevins et al. [66, 65] found that signals in the theta range in the frontal midline regions of the scalp were greater with tasks demanding increased levels of working memory. This is in line with other studies which have found increased amplitude in theta EEG signals in the frontal midline regions (e.g., [75]). Another finding has been a decrease in the alpha EEG band by tasks that increase cognitive workload, e.g., [66, 65, 94]. Interestingly, alpha signals increased after practice on the tasks, suggesting that this signal is inversely related to the amount of cortical resources devoted to the cognitive task [66].

Another use of EEG signals to measure cognitive workload is by looking at the P300 event-related potentials (ERP) signal. It has been shown that the more cognitive effort is put into a task, the higher the amplitude of the P300 signal [196]. However, in order to garner this measurement, a secondary task is required to be measured concurrently with the primary task. This has the disadvantages discussed in section 2.2.3.2. Ullsperger et al. [210] looked at the P300 that responds to novel auditory stimuli whilst carrying out two cognitive tasks. They found decreased amplitudes of the P300 in response to novel sounds when cognitive workload was high. Similar results were found in another study using auditory tones while users were carrying out cognitive tasks [187]. However, once again this method requires the use of secondary stimuli which can take away precious, finite cognitive resources from the primary task, especially if the cognitive task at hand is particularly complex with high element interactivity.

Szafir and Mutlu [207] monitored student attention with EEG and determined which lecture topic students would benefit from reviewing. Szafir and Mutlu [206] used EEG to detect and recapture dropping attention levels in real-time. They found this improved recall abilities. This is a rare example of how real-time adaptations can help learning using physiological sensing. Instead of measuring attention levels, however, in this thesis I will be measuring user cognitive workload. The amplitude of the P300 signal is also increased by changes in cognitive workload [196],

however this method requires the use of secondary stimuli which can take away precious, finite cognitive resources from the primary task.

MRI and PET studies have established that cognitive workload can be measured in the prefrontal cortex[36, 121]. More recently, brain imaging studies have examined changes in brain function associated with improvements in performance of tasks [63, 15, 171, 92]. Activations seen earlier in practice such as the prefrontal cortex, decrease in activation later on, while other areas associated with highly practised performance in task-specific areas, increase in activation with task practice [92].

There has been some disagreement regarding the effects of practice on functional neuroanatomy, with effects including increase or decrease of activation, as well as changes in functional distribution (see [92] for a review). Most neuroimaging studies have found decreases in functional activation with task practice, e.g., [15, 63]. For example, PET imaging of cognitive skill learning in the brain showed that early performance of a task was associated with increased blood flow in certain areas including the dorsolateral prefrontal cortex and orbitofrontal cortex [15]. Improved performance on the task was associated with a significant decrease in areas of the prefrontal cortex over the course of learning. This is thought to be due to increased neural efficiency so that a minority of neurons fire strongly during the task [171], thereby making practice-related activation decreases the result of more efficient use of neuronal firing [92]. Activation increases have also been found, however, this may be due to practice-related expanded representations [171]. It has also been suggested that there are 'scaffolding-storage' areas that are used at the start of novel tasks, and that with practice, areas in the brain that can more efficiently store and access information are used, allowing the scaffolding network to fall away [162, 92]. This makes activations seen earlier in practice such as the prefrontal cortex, decrease in activation later on, while other areas associated with highly practised performance in task-specific areas, increase in activation with task practice [92].

2.3 Creativity

As background literature for Experiment 3, Chapter 5, I shall discuss creativity in the context of cognitive workload and adaptive, intelligent systems.

2.3.1 Creativity in HCI

Creativity in HCI has been a greatly discussed area. Shneiderman presented the "*grand challenge for creativity support tools*" for HCI researchers [190]. An NSF Workshop was carried out on the need for creativity support tools in 2005 [191, 192]. At the workshop, creativity was loosely defined on an individual level as involving "*the development of something novel and valuable to the individual*" (page 11) [191]. It was also recognized that there are multiple dimensions to creativity, and therefore to a taxonomy for creativity studies and creativity support tools. For example, if one was to work with individuals as opposed to groups, such as I do in this thesis, the creativity support tools should be specifically designed for the correct intersection of the n-dimensional taxonomy [191, 192].

The workshop self-reportedly "*relied heavily*" upon Mayer's (1999) book [122]. In his book, Mayer discusses the fact that physiological measures are becoming available to the study of creativity, and points out that *"brain event recording data provides information not available to other methodologies*". This is precisely where the next generation of brain-computer interfaces can be of service to the field of creativity: by providing internal readings of users' brain state, systems can be built that detect and respond intelligently to aid creativity.

Yet, as pointed out by Mayer [122] and echoed by Shneiderman et al. [191, 192], it is not yet clear *how* such a system could provide creative support to a user. And that is precisely the problem I address in Chapter 5. I tackle this problem using the medium of music and thus, in the next section, I discuss musical creativity and brain-computer interfaces.

2.3.2 Musical Creativity and BCIs

Traditionally, most musical BCI systems have been based on a mapping of brainwaves to audio signals. One of the first musical BCIs was reported in the seminal paper by Adrian and Mathews 1934 who translated electroencephalogram (EEG) signals into sound [2]. In 1965, the first musical piece was performed with a BCI by directly mapping the players alpha brainwaves over loudspeakers to resonate onto percussion instruments [114]. In their biosignal musical interface, Knapp et al. used EEG alpha waves in the occipital cortex to change the MIDI program [95]. In their demonstration they gave they example of changing from a violin to a glockenspiel sound [95]. The applications of this system for disabled people was discussed in Scientific American [116]. More recently, Miranda et al. built a BCI-Piano that responded to certain frequencies of EEG activity with assigned musical passages [138, 139]. Arslan et al. used a combination of eyeblinks based on EEG alpha-bands and motor imagery stimulation from the motor cortex to apply a mapping to note triggers and diffusion of sound over loudspeakers [10]. Mealla et al. used the magnitude spectrum of EEG to shape the spectrum of white noise in their musical table-top interface [130].

There have also been examples of BCIs where direct brain signals are used to control a grid of options to select icons as a replacement for a keyboard and mouse in order to compose music using the P300 signal [27] and steady-state visually evoked potentials [104]. While such systems are excellent designs for users with motor disabilities, they are slower to operate than a mouse and keyboard for able-bodied users.

Musical BCIs have also used brain signals as direct input to an application, such as using the P300 signal to directly control a synthesiser [72] or select harmonies [211]. Icons from a grid have been selected to compose music instead of the standard keyboard and mouse using the P300 signal [27] or steady-state visually evoked potentials [104]. While such active BCIs can be invaluable to people with motor disabilities, they are too slow and inaccurate for the general population

compared to the standard mouse and keyboard.

Recently, however, advances in brain-sensing technologies have been used for implicitly detecting and responding to users where the user performs a task normally, but the user's brain signals act as an *additional* communication channel that are used to assess cognitive state and adapt the interface accordingly. Implicit BCIs are characterized by implicitly by measuring user cognitive state without any additional effort from the user [235].

A clear need for higher level semantically meaningful brain data has been highlighted in the field of musical BCIs [138, 141]. Rosenboom [180] discussed the possibility of using Event-Related Potentials with EEG.

Miranda et al. [141] ran a pilot study on a EEG-based BCI that detected active versus passive listening. Active listening is when a user would imagine the riff continuing for 2 bars after it finished whereas passive listening involved no extra effort. However, they reported difficulty in reliably detecting listening behavior [140]. They also discussed building another musical BCI that would direct the tonality of the music based on auditory cortex stimulation. However they reported that only the generative music system had been built so far which is based on hypothetical brain data [140]. Girouard et al. used fNIRS to classify between two non-music related tasks and changed the background music according to the predicted task [67]. They found no effects of background music on the tasks but their main goal was to produce a proof-of-concept passive adaptive BCI rather than a musical BCI [67]. Grierson et al. [73] have used the Neurosky headset's attention and meditation levels to control the systems Brainemin and Brain Controlled Arpeggiator in live performances. However there has been no user evaluation or controlled experiment using these cognitive states.

The time is therefore ripe to create a novel musical interface to aid creativity that *implicitly* senses and responds to the user's cognitive state.

The system, BrAAHMS, discussed in Chapter 5 addresses both of the issues raised above. It derives *higher level semantically meaningful brain data* from users, and it does so in an *implicit* manner.

2.4 Trust

As background literature for Experiment 4, Chapter 6, I shall now discuss the elements of user trust in intelligent, adaptive systems.

2.4.1 Cognitive and Affective Trust

Trust has been defined as *"an evolving, affective state including both cognitive and affective elements and emerges from the perceptions of competence and a positive, caring motivation in the relationship partner to be trusted"* [229]. What is interesting about this definition is its intrinsic ties with user affect. This idea of trust containing both affective and cognitive elements was first introduced in the field of sociology [112]. In the context of user-agent interactions, *affective trust* is related to the perception that an agent's actions are intrinsically motivated and is based on affect generated by care, concern, and security [86, 176]. *Cognitive trust* is based on rational thought processes and empirical evidence [112]. Cognitive trust is very closely related to an intelligent agent's competence as defined by [118], whereas affective trust could be developed by ensuring and enhancing other factors in the agent.

Dowell et al. [45] found that in the early phase of a relationship cycle in business (in the first 12 months), affective trust was especially important. In later, more mature phases of the relationship cycle (after the first 12 months), cognitive trust was a more important factor than before [45].

When there is less substantive evidence to build trust upon, it is thought that trust is based upon the hope or expectation that the task will be fulfilled, therefore cognitive trust would be less influential early on [45]. Lewis and Weigert [112] suggest that the stronger the affective trust in relation to the cognitive trust, the less likely behavioral evidence to the contrary will weaken the relationship.

Affective trust has been modeled into two sub-categories by Dowell et al. [45]:

- *Relational trust*, which is related to the 'leap of faith' aspect of affective trust

where faith is placed in the other individual that they will act in a trustworthy way.

- *Intuitive trust* which is based on moods and feelings about the other individual.

Dowell et al. [45] found that relational trust was particularly important in the early phase of a relationship.

2.4.2 Trust between Humans and Computers

Research on trust between humans and intelligent systems can be divided into two broad categories: 1) *The algorithms of the decision making* - i.e., the *what* and *when* of the system's adaptations, and 2) *The interaction style and design* - i.e., *how* the system interacts with the user.

The algorithmic decision making, i.e., the artificial intelligence of the system, is fallible with current technology. While errors still exist in intelligent systems, as HCI researchers, we can draw from the power of the second category, the interaction style and design, to cope with deficiencies in the machine. We discuss the two categories in further detail below.

2.4.2.1 Trust and the Decision Making Algorithms

The importance of the adaptations made by the system can determine whether the intelligent system should be entrusted to intervene and what actions it should take [31]. Transparency of such algorithms behind the system's decisions, knowledge of the system's resources, [69] and inclusion of comprehensive product information [107] have been found to be important to engendering trust with the user. Not only do users want to understand the decision making processes, but they also would like mechanisms to override erroneous behavior when necessary [69]. They also are much more reluctant to trust intelligent systems to make important changes such as making a purchase with a credit card [69].

System reliability is crucial to building trust between the human and computer and needs to be built early on [120, 110]. Maltz and Meyer [120] provided

users carrying out a visual task with potentially helpful cues that varied in their reliability. They found that only the users who received highly reliable cues in the first set of trials continued to accept suggestions in the second set of trials. LeeTiernan et al. [110] found that when users' first impression of system reliability was high, they continued to use system suggestions, even when reliability decreased. Conversely, when first impressions denoted low system reliability, users would not accept system suggestions even when reliability increased [110]. Similarly, when users viewed an algorithmic error in a forecasting system, they lost confidence in the system and chose an inferior human forecaster over it [37].

2.4.2.2 Trust and the Interaction Style and Design

HCI researchers have been taking paradigms in social psychology used in human to human relationships and applying them to develop human to computer relationships [175, 18]. Transfer of results has been found using these paradigms as people do demonstrate anthropomorphic behavior towards computers [146]. Behaviors include applying human social categories such as gender and ethnicity to computers or engaging in social behaviors such as politeness and reciprocity [175, 146].

Social psychology has shown that humans are persuaded through association with good feelings [62, 188, 117] and that they like people more who like them [93]. So it is not surprising that flattery can be a powerful tool to increase likability [16].

Fogg and Nass [54] found that humans are susceptible to flattery from computers and that the effects are the same as flattery from humans. Participants subjected to flattery reported greater performance and more positive evaluations of the human-computer interaction [54]. Lee [105] also found flattery increased positive impressions of the computer, however, flattery was also found to increase user suspicion about the validity of computer feedback and actually lowered acceptance of computer suggestions. A suggested explanation for this has been that flattery temporarily fosters mindfulness, thus making users scrutinize the available information [106].

The similarity or familiarity paradigm in social psychology has demonstrated

that people will like others who are similar to them [74, 91]. Users prefer computers that match them in personality over those that do not [175]. Interestingly, users actually prefer computers that become more like them over time more than computers that maintain a consistent level of similarity [175]. Users also put more trust in systems that are more human-like [216].

Other relational behaviors used in human relationships have been adopted by HCI researchers. Virtual agents that use humor are rated as being more likable, competent and cooperative [145]. Computers that engage in reciprocal and deepening self-disclosure during conversation engender more personal information from the user and increase chances of a product sale [143]. Highly expressive pedagogical agents increase students' perception of trust [111]. Bickmore and Picard [18] found that a relational embodied agent that used social dialogue, meta-relational dialogue, politeness, non-verbal empathy exchanges, and continuity behaviors was trusted, respected and liked more than a non-relational agent.

The physical attractiveness stereotype is a long-held paradigm in social psychology stemming from "what is beautiful is good", whereby people attribute other positive traits onto attractive people [39]. I will now discuss it in further detail.

2.4.3 Physical Attractiveness Stereotype

The unfortunate and uncomfortable truth is, that as human beings, we are positively biased towards physically attractive people. Physically attractive people are thought to possess more favorable characteristics such as being more sociable, happier, sexually warmer [135, 39, 5], better adjusted and more self-assertive [47], more intelligent [83], more successful in life [39, 148], be better marital partners [148], have greater social power [136, 195], have more positive effects on other people and receive more positive responses from others including requests for help and requests at work [128], more persuasive [14, 42, 23], and more qualified for jobs [79] than less physically attractive people.

The physical attractiveness stereotype has held up consistently across different populations, and exists in older, more sophisticated groups such as health pro-

fessionals [150], teachers [49] and other professional groups [5]. The stereotype also holds in young children who are biased towards more attractive children [40, 41, 103]. Even 3-month-old infants show selective preference for attractive over unattractive adult faces [102, 184], suggesting that attractiveness standards are more inherently acquired.

While two reviews have found the physical attractiveness stereotype is stronger for certain personality traits, with less correlation across integrity of character and attractiveness [47, 53], there is a wealth of evidence across many fields to suggest that there is certainly some favorable bias in the area of trustworthiness and physical attractiveness. Patzer [157] found that communicators with higher levels of physical attractiveness were perceived as more trustworthy and of higher expertise than communicators of lower physical attractiveness. Higher communicator physical attractiveness also had a significant effect for increased liking of the communicator [157]. This ties in with findings that celebrities who are liked will also be trusted [59]. Friedman and Friedman [58] found a significant correlation between physical attractiveness and trustworthiness of political figures.

This correlation between physical attractiveness and trustworthiness carries into breaches of trust as well. Adults view more attractive children's transgressions less negatively. Moreover, when the child is unattractive, the offense is seen to be more of an enduring dispositional fault [38].

This bias translates into later life where mock juror studies show that physically attractive defendants are treated more leniently than unattractive defendants for many types of crimes such as robbery, rape, cheating, and negligent homicide [48, 194, 124]. Physically attractive defendants are found to be guilty less often and are given shorter sentences [48, 124]. This has not been found to be the case for the crime of swindling where the crime is attractiveness-related [194, 124]. While Mazzella and Feingold [124] suggest that such results are linked to defendant likability in that attractiveness may be promoting liking, Abwender and Hough [1] controlled for likability and found unchanged results, suggesting that jurors are responding to something more than just likability.

The correlation between defendant physical attractiveness is also externally valid as corroborated by length of sentencing in real courtrooms by judges and juries. Stewart [201] found that attractiveness was highly predictive of both minimum and maximum sentences, with the more attractive the defendant, the less serious the sentence imposed. These findings were confirmed by Stewart [202] and were found to remain significantly correlated even when the seriousness of the crime was controlled. Downs and Lyons [46] investigated judges' sentencing of 915 female and 1,320 male defendants who were unable to alter their appearance levels prior to court appearances. They found, once again, that judges exhibited a strong attractiveness-bias in the bails and fines that they set (for misdemeanor charges) [46].

Chapter 3

Experiment 1: Feasibility Study

In order to establish whether varying levels of cognitive workload can be measured while users play piano on a musical keyboard, I first ran a feasibility study. This study was the basis of Experiments 2 and 3.

3.1 fNIRS System

3.1.1 Equipment and Measurements

A multichannel frequency domain Imagent fNIRS device from ISS Inc. (Champaign, IL) was used for data acquisition. Two probes were placed on a participant's forehead to measure data from the two hemispheres of the prefrontal cortex. Each probe contained four light sources, each emitting near-infrared light at two wavelengths (690 and 830 nm) and one detector; thus there were sixteen data channels (2 probes x 4 source-detector pairs x 2 wavelengths) (Figure 3.1). The source-detector distances ranged from 1.5 and 3.5 cm, and the sampling rate was 11.79 Hz.

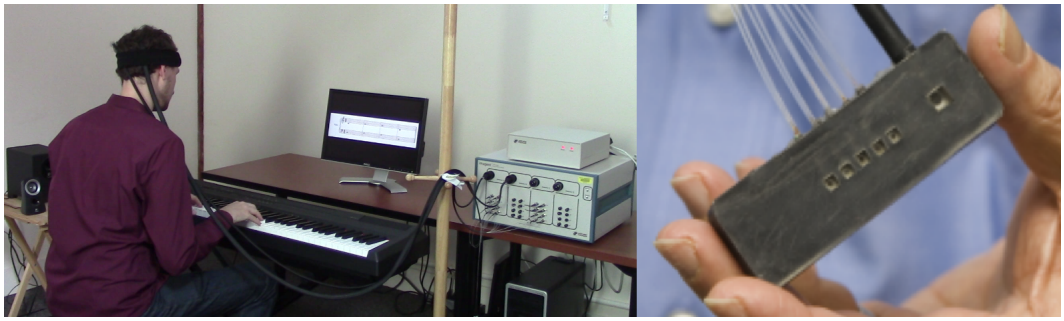


Figure 3.1: fNIRS equipment and experiment setup. Left: Subject is playing the piano keyboard while wearing the fNIRS sensor. Imagent is visible on the right. Right: An fNIRS sensor with light sources (left) and one detector (right).

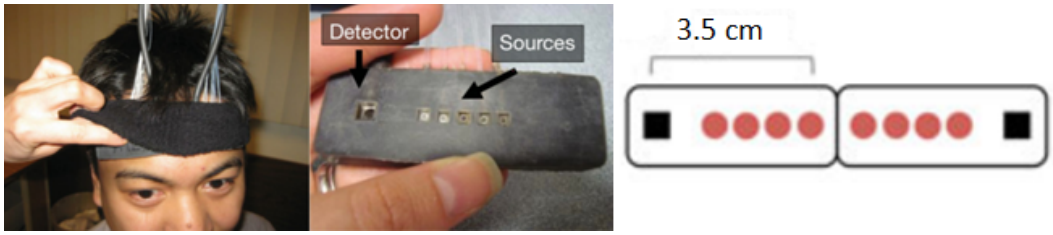


Figure 3.2: Each fNIRS probe had 4 sources (five are shown but four were used) and one detector which were placed on the forehead to take measurements from the prefrontal cortex [161].

3.1.2 Procedures

The feasibility study was carried out in order to determine whether high and low cognitive workload could be classified using fNIRS while participants played musical pieces on the piano keyboard. The training task for the brain data consisted of participants playing 15 easy and 15 hard pieces on the piano for 30 seconds at a time. During each musical piece, the system calculated the change in optical intensity compared to a baseline measurement for each of the sixteen channels. Markers sent at the beginning and end of each trial denoted the segments for each piece.

3.1.2.1 Training Task and Materials

Previous fNIRS studies have classically used the n-back task to elicit high and low cognitive workload (e.g. [160, 3]). Users are shown varying squares on a screen for 500ms at a time and are asked to indicate whether the current stimulus is the same as a stimulus n squares back. There are 2 conditions: a 1-back condition to elicit low cognitive workload and a 3-back to induce high cognitive workload (Figure 3.3). Users typically carry out 15 trials of each condition with 10 stimuli in each trial. Each trial totals 25 seconds and is followed by a 15 second break.

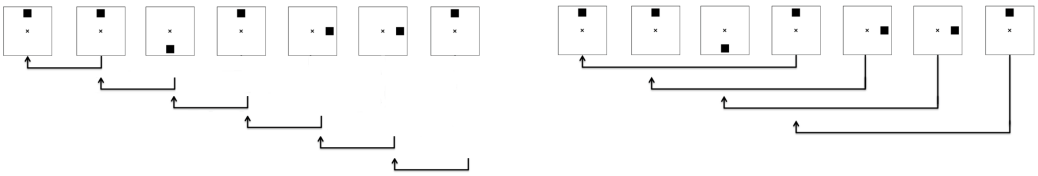


Figure 3.3: N-back task: Each square is flashed on the screen one at a time. Left: 1-back elicits significantly lower cognitive workload than Right: 3-back which requires holding more information in the short-term memory.

In this thesis, I chose an equivalent musical task to elicit states of high and low cognitive workload. I chose 15 easy pieces and 15 hard pieces for participants to play on the piano for 30 seconds at a time. The criteria for the easy pieces were that a) all notes were in C major (i.e. no sharps (♯) or flats (♭)), b) there were only whole notes (𝅝) (very slow, long notes), c) there were no accidentals (i.e. no additional sharps, flats, or naturals (♮) that are not part of the scale), d) all notes were within the C to G range so that participants did not need to move their hands e) there were no dynamics (i.e. volume of a note or stylistic execution). The hard pieces were chosen by a musicologist and the criteria consisted of pieces that a) had a harder key signature (most pieces had a key signature of at least 3 sharps or flats), b) contained accidentals , c) contained mostly eighth (♪) and sixteenth notes (𝅘𝅥𝅯) (i.e. short, fast notes), d) required some moving of the hands but not too excessively, and e) contained dynamics. Figure 3.6 shows an example of an easy and hard piece.

All pieces can be viewed in Appendix A.

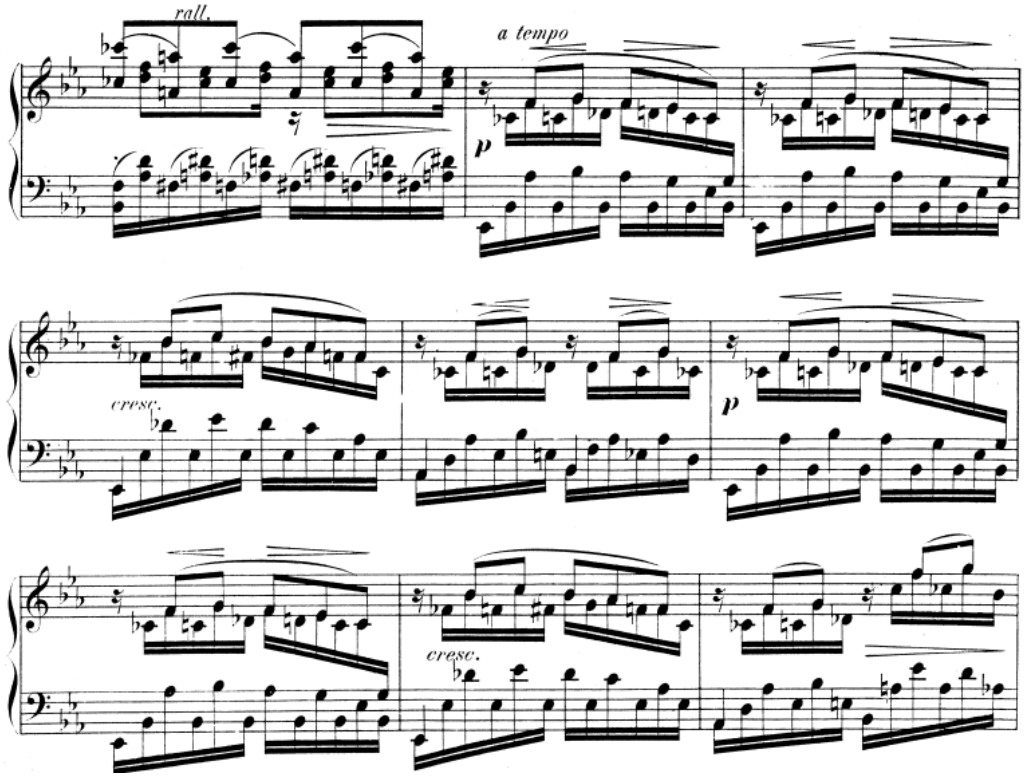


Figure 3.4: An example of a hard piece, excerpt from Impromptu No.1, Opus 25 by Gabriel Fauré

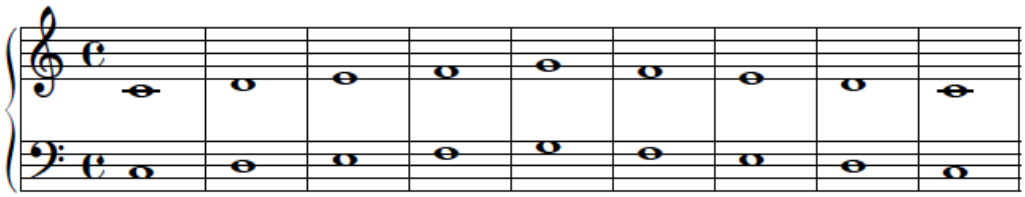


Figure 3.5: An example of an easy piece

Figure 3.6: An example of a a) hard piece and b) easy piece. Both pieces are nine measures long.

Although it seems at first glance that the two different musical scores in Figure 3.6 would cause a great difference in hand motion, participants actually only managed to play one or two measures from the hard pieces very slowly. This is

because the hard pieces were purposefully chosen by a musicologist to be complex. Furthermore, sight-reading (playing a piece that you have never seen before) is a challenging task by itself. Therefore the difference in hand motion between the easy and hard pieces was minimal. The fNIRS signal is resilient to minor hand movement [197] as caused by the simplicity of the easy pieces and the complexity of the hard pieces.

3.1.2.2 Experimental Procedure

Participants were given 15 easy and 15 hard pieces of music in random order to play on the piano. They were given 30 seconds to sight-read each piece (i.e. play a previously unseen piece) followed by a 30 second rest period. A metronome was played at the start of each piece for 4 seconds at a speed of 60 beats per minute. Participants were asked to try to stick to this speed but told they could go slower if they needed to. Figure 3.1 shows the setup of Experiment 1.

3.1.2.3 Participants

Fifteen participants took part in the first experiment (7 female, mean age of 21, SD of 3.2) and were compensated $15 for participating. Subjects had been playing piano for a mean of 9 years (SD 5.4).

3.1.3 Data Analysis and Classification of fNIRS Signal

Figure 3.7 shows the overall signal analysis and classification of the fNIRS system. In this section I will be discussing each step in detail, with further details for the real-time system in Chapters 4 and 5.

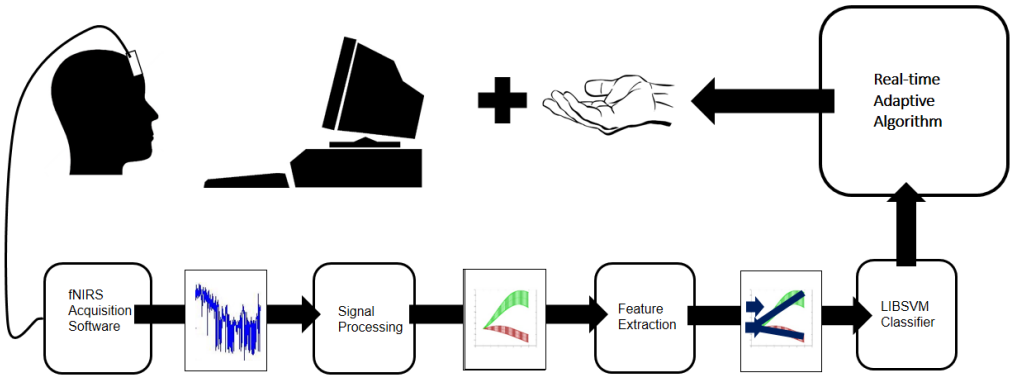


Figure 3.7: Overall signal analysis and classification of the fNIRS system.

3.1.3.1 Signal Preprocessing

The optical intensity signals were filtered for heart rate, respiration, and movement artifacts using a third-degree polynomial filter and low-pass elliptical filter which has been standard in HCI fNIRS studies (e.g., [78, 198, 199, 3, 4]).

3.1.3.2 Feature Extraction

Once the fNIRS signal has been pre-processed, it must then be classified into states of high or low cognitive workload. Feature extraction is important to help avoid overfitting in machine learning. As part of my thesis, I implemented feature extraction in the analysis of the brain signals.

There have been a range of features used to classify the fNIRS signal. Solovey et al. [197, 198, 199] used each time point from each channel over the entire trials as individual features. Hirshfield et al. [78] used the maximum, minimum, and mean of the concentration signals as well as the slope, time to peak, and full width at half maximum as their features. Moghimi et al. [142] used two genres of features. Their single-channel features were the mean, linear regression slope of the trial window

and coefficient of variation (ratio of the variance to the mean) of the concentration signals. The two laterality features used were the ratio of the concentration signal slopes and the difference in the mean signal values [142]. Afergan et al. [3, 4] used the mean linear regression slope of the trial window as the individual features.

In this thesis, I will be using the most commonly used features, the mean and linear regression slope, as the extracted features that will be input into the classifier. The mean is calculated from each channel's 30-second trial window. The linear regression slope is calculated by fitting a line using linear regression to all data points in each channel's 30-second trial window and calculating the corresponding slope. Figure 3.8 illustrates an example of the features from one channel from one individual averaged across 30 trials of 30-second trial windows. The mean and linear regression slope can be seen highlighted for both the easy and hard pieces.

As there 16 channels in total, there are 32 features in total (16 information channels x 2 descriptive features).

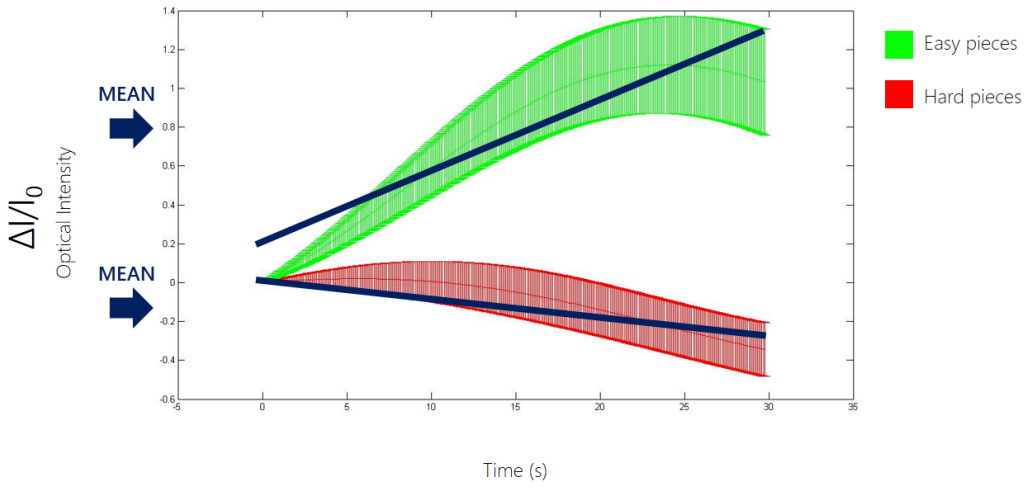


Figure 3.8: Relative change in optical intensity of one individual averaged across 30 trials for 30 seconds per trial for one channel. The two features, mean and linear regression slope, are highlighted.

3.1.3.3 Feature Selection

The classification algorithm chosen, which I discuss in detail in the next section, selected the most ideal features to use out of the total 32 features discussed above for each individual.

Figure 3.9 contains one individual's brain data averaged across 30 trials for 30 seconds per trial across all 16 channels. But most importantly, figure 3.9 shows that the machine learning algorithm discussed in the next section selected certain features from certain channels. The features selected from various channels were different for each individual.

3.1.3.4 Classification

Machine Learning Algorithm

The features selected and the classification technique are thought to be dependent on each other [25] as well as being dependent on the task. Similar to feature extraction, there have been a variety of classifiers used in fNIRS research. Solovey et al. [198, 199] implemented the sequential minimal optimization (SMO) algorithm [170] for support vector machines (SVMs) to classify across three multitasking states. Hirshfield et al. [78] used a Naive Bayes classifer to classify between high and low cognitive workload. Girouard et al. [68] used a k-nearest neighbor (kNN) algorithm to classify users' cognitive responses between easy and hard game difficulty levels. Moghimi et al. [142] used linear discriminant analysis (LDA) to solve two different two-class problems of arousal and valence to measure emotion.

Afergan et al. [3, 4] implemented the SMO algorithm for SVMs using the open source machine learning library LIBSVM [26] to classify multitasking and cognitive workload.

In this thesis, I also used LIBSVM to implement the SMO algorithm for SVMs to classify between high and low cognitive workload.

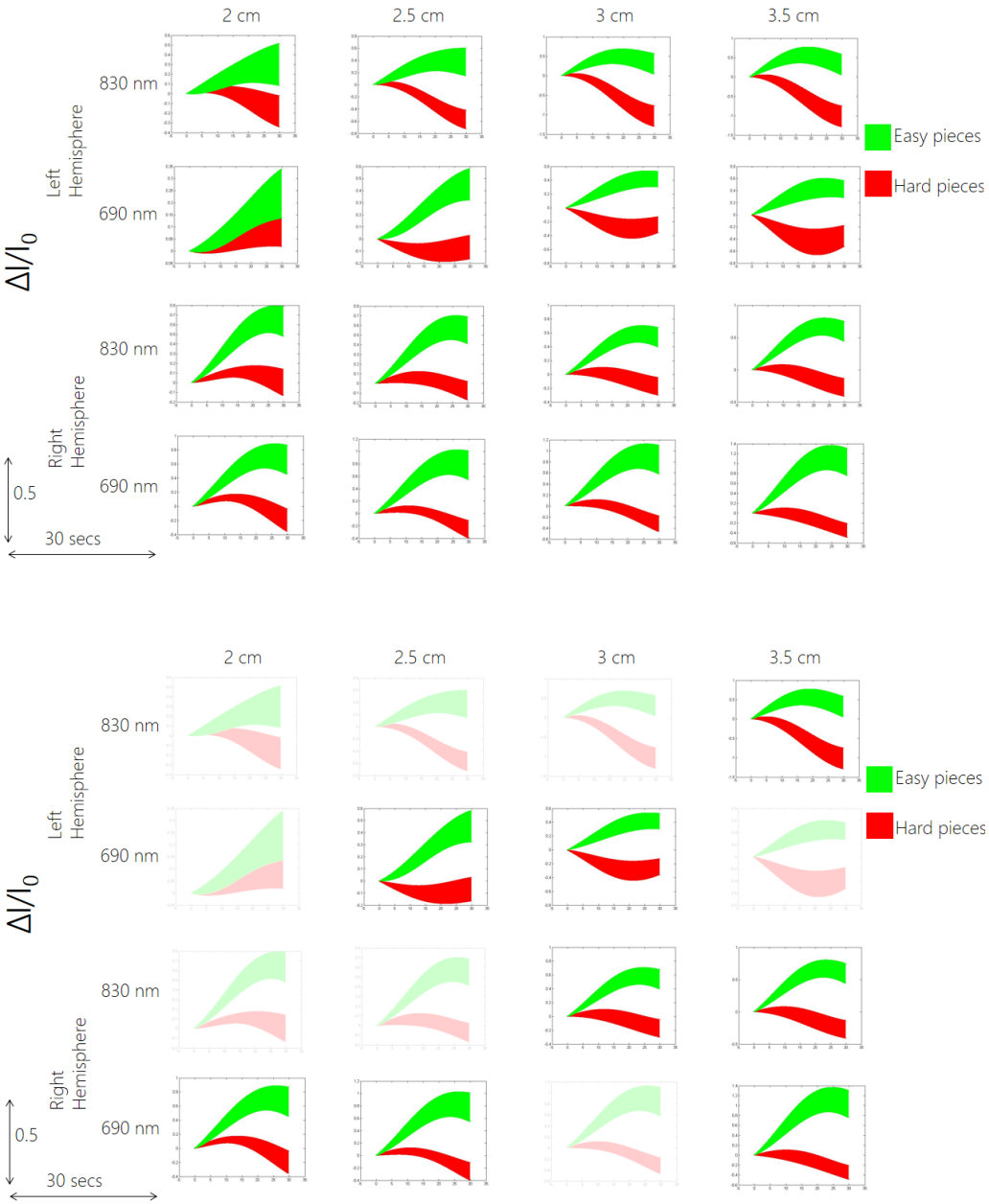


Figure 3.9: Example of Top: Relative change in optical intensity of one individual averaged across 30 trials for 30 seconds per trial across all 16 channels. Bottom: The highlighted channels are the ones used to pick features by the LIBSVM machine learning algorithm. The algorithm picked different channels for different individuals.

Classification Procedures

For each training trial, a 30-second sliding window of data was extracted of the time series of change in optical intensity compared to an initial baseline period of

30 seconds for each of the 16 information channels (2 probes x 4 light sources x 2 wavelengths). Each of the 32 features described above were fed into the LIBSVM classifier to solve the two-class cognitive workload classifications.

Cui et al. [25] compared linear and nonlinear SVM performance in NIRS data classification and found that nonlinear SVM tended to overfit. I therefore used the linear kernel function within LIBSVM.

Amongst LIBVSM parameters are: $C > 0$, which is the penalty parameter of the error term, and γ, which is one of the kernel parameters for nonlinear kernels. As I was using linear SVM, there was only free parameter (C). In order to find the optimal value of C I carried out a parameter search for each individual to train the whole training set. Finally, after the machine learning model was built on the training data, LIBSVM carried out a 10-fold cross validation to ensure that the model was accurate and to construct a probability model of estimates for the test data.

Probability Estimates of Real-Time Data

The probability estimates of model are a function of LIBSVM that can be turned on and they are values between 0 and 1 of how likely the test data belongs to one class or another. I used them in order to get a more accurate idea of the test data. While the probability estimate is not simply the distance between the separating hyperplane and vector, it does take this distance into account. For example, if the distance is very large, the model can have a high degree of confidence that it was correctly classified, thus the probability estimate with be very high, either at or close to 1.0. In contrast, if the vector is very close to the hyperplane separating the two classes, the probability estimate will be close to 0.5 for each class.

The probability estimates in LIBSVM are for binary classes, so if the vector is in class high cognitive workload (HCW), we compute $P(HCW) \in [0, 1]$ and then the probability of not high cognitive workload is $P(HCW^c) = 1 - P(HCW)$.

The probability estimates can be used with real-time data as well as to classify training data. In this chapter, offline analysis was used to classify between high and low cognitive workload during the feasibility study. In Chapters 4 and 5, I

will discuss real-time analysis of the brain data that allows for adaptive interfaces.

3.2 Experiment 1 Results and Discussion

Figure 3.10 shows the mean and standard error in the oxygenated hemoglobin of participants while they played easy (blue) versus hard (green) pieces on the piano. Although we built an individual model for each participant, we present the mean findings across all 15 participants across all 30 trials in Figure 3.10 to illustrate this general trend.

To investigate differences between hard and easy pieces, we performed a t-test on the mean change in oxygenated hemoglobin. This revealed a significant difference between conditions when participants played an easy piece ($\mu = -0.1, \sigma = 0.1$) versus a hard piece ($\mu = -0.02, \sigma = 0.1$) on the piano ($t(14) = -3.04, p < .01$). Means and standard errors are shown in Figure 3.10.

The significantly higher levels of oxy-Hb when participants were playing harder pieces on the piano correspond with the hemodynamic literature, whereby, when there is increased cognitive activity in an area of the brain, excess oxygen is provided to that area. The increase in oxygen consumption is less than the volume of oxy-Hb provided, resulting in more oxy-Hb [55].

As discussed in the previous section, due to the simplicity of the easy pieces which required very little hand movement, and the complexity of the hard pieces, where participants could not get further than the first or second measure very slowly, participants didn't move their hands very much. In light of the resilience of fNIRS to minor hand movement [197], the significant differences in oxy-Hb reported in Figure 3.10 can be related to differences in cognitive workload in users when playing easy versus hard pieces on the piano.

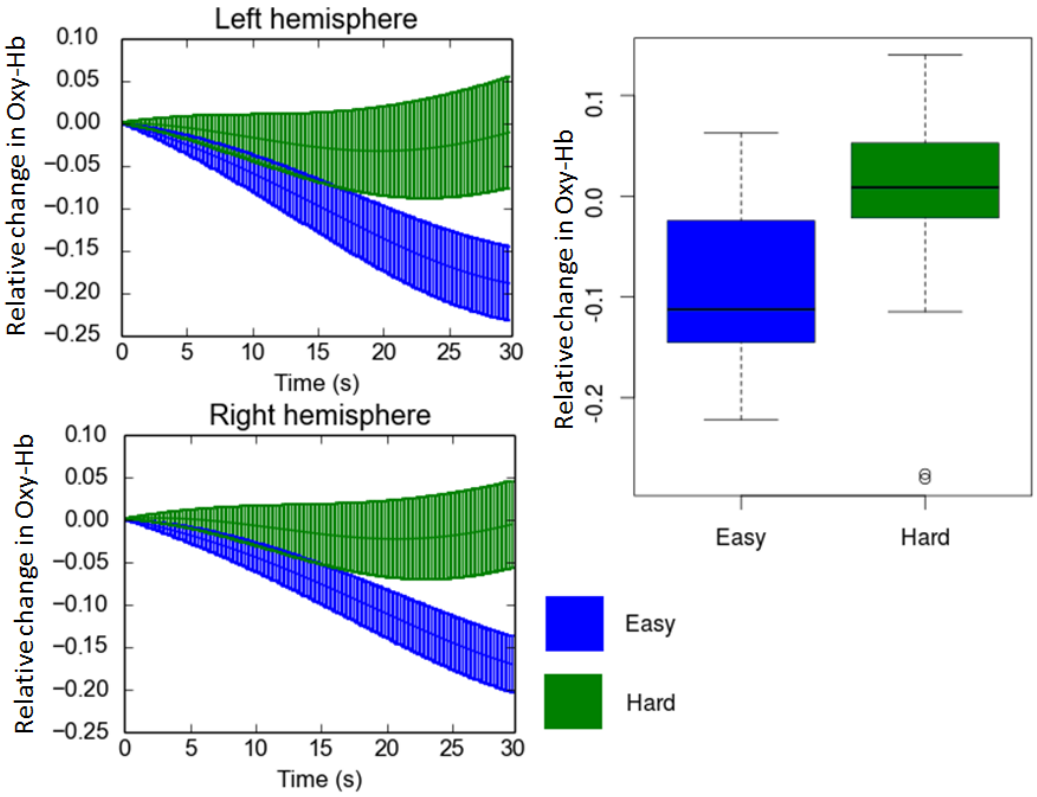


Figure 3.10: Left: Mean change and standard error in oxy-Hb in Experiment 1 across all participants. Although each participant was modeled individually, the fNIRS signal exhibited a general trend with higher levels of oxy-Hb corresponding with hard pieces. Right: The mean change in oxy-Hb was significantly higher in participants when they played a hard piece than an easy piece ($p < .01$).

3.3 Experiment 1 Conclusions

Experiment 1 established that there were differences in the fNIRS signal when participants played easy vs. hard pieces on the piano. I therefore built upon these findings to design and create two novel musical brain-computer interfaces: BACh: Brain Automated Chorales, and BRAAHMS: BRain Automated Adaptive Harmonies in a Musical System which will be discussed in Chapters 4 and 5 respectively.

Chapter 4

Experiment 2: Learning

This work won a Best Paper Award at ACM CHI 2016. Yuksel et al. (2016) Learn Piano with BACh: An Adaptive Learning Interface that Adjusts Task Difficulty based on Brain State. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems.* pp 5372–5384 [233].

4.1 Experiment 2 - Introduction

Good teachers, whether they realize it or not, often guide their students into the *zone of proximal development*. That is, they help their students fulfill a learning potential with guidance that the students could not have reached by themselves [214]. However, this is not always an easy task as learning is a complex activity affected by factors such as difficulty of the task to be learned, learner cognitive ability, instructional design, and learner motivation.

Intelligent tutoring systems (ITS) and other computer-based education (CBE) systems have been developed in the last decades to aid this learning process. There have been promising results; meta-analyses revealed that students tend to learn faster and more accurately, and have increased enjoyment of their learning material [100].

Cognitive learning theory (CLT) plays an important role in the designing of instructional learning material and systems. It is based on idea that there is only

a finite capacity for working memory available [177, 204] so designers of learning material should avoid overloading students. One approach has been to present learners with tasks of increasing difficulty. Interestingly, meta-analyses have revealed that this is only helpful if the levels of difficulty are increased *adaptively* to learners [219, 220].

Adaptive, interactive learning tasks generally tend to use a mixture of performance and cognitive workload of learners measured by self-reporting on a one-item scale (e.g. [183]). While these studies have made progress one factor is repeatedly highlighted as the weak link in CLT studies and learning studies in general: the measurement of cognitive workload [34, 144, 90, 20]. An adaptive learning system that could adjust task difficulty based on the learner's actual cognitive state, rather than their task performance, would address this issue.

Brain sensing using functional near-infrared spectroscopy (fNIRS) has recently been demonstrated by HCI researchers to measure user cognitive workload in adaptive systems [12, 199, 3] and is resilient to subject movement as might be seen while playing a piano [197]. In this thesis, I use brain sensing as the *determining factor* in a learning task by increasing task difficulty adaptively when the learner's cognitive workload falls below a certain threshold, demonstrating that they can handle more information.

I present **Brain Automated Chorales (BACh)**, an adaptive musical learning system that helps users learn to play Bach chorales on the piano by increasing the difficulty of the practice sessions based on cognitive workload using brain sensing. BACh aims to guide learners into the zone of proximal development by measuring when they can cognitively handle more information and providing them with the next stage of learning at the right moment.

I use a musical task on the piano keyboard because it lends itself well to task segmentation with increasing levels of difficulty and to high element interactivity due to concurrency. A musical task is also easy to evaluate in terms of accuracy and speed when compared with a control condition. As demonstrated in Chapter 3, previous work has shown that cognitive workload can indeed be measured with

fNIRS brain sensing while playing the piano [230].

By analyzing performance data, we found that participants played the pieces faster and with increased accuracy when using our BACh system compared to how they would normally learn a piece of music. Participants also subjectively reported that they played the pieces better, and liked the timings of the level changes.

Thus, the contributions of this chapter are as follows:

1. Build an adaptive, interactive musical learning system, BACh, that guides learners' progress by measuring when they can cognitively handle more information and providing them with the next stage of learning at the right moment.

2. Demonstrate that users could play the musical pieces faster and with more accuracy using BACh than they normally would in a control condition.

3. Demonstrate through questionnaire and interview data that participants felt they learnt the pieces better with BACh and they liked timings of the level changes.

4.2 Experimental Design

Sixteen participants (8 female, mean age of 21, SD of 2.4) took part in a within-subject design. All participants first undertook a *training task* where they played 15 easy and 15 hard pieces on the piano. Each piece was 30 seconds long with a 30-second rest between each piece.

After completion of the training task participants received a break and then moved onto the *learning task*. During the learning task, each participant learned two Bach chorales. They had 15 minutes to learn each chorale. One chorale was presented in normal form and participants were instructed to learn the piece the way they normally would. The other chorale was presented with our adaptive interface BACh. The order of the conditions alternated between participants. At the end of the allotted learning time, participants were asked to play the piece once all the

way through the best they could. Performance data from both conditions were recorded and evaluated based on the performance where they played the piece all the way through as best they could. Figure 3.1 shows the fNIRS equipment and experimental setup. Participants wore the fNIRS equipment throughout the entire experiment, including both the control and experimental conditions.

At the conclusion of the learning task, participants were given a questionnaire on how they felt they had learned the two pieces. We also gave participants a short interview on what they thought of the adaptations made by BACh.

All participants were compensated $20. While eleven out of 16 participants rated themselves as beginners, out of the remaining five who rated themselves as intermediate, 3 of them no longer played piano. The median time participants had played piano was 3 years.

The details of the materials, tasks, and data analysis are given below.

4.3 Learning Task

4.3.1 Music for Learning Task

For the learning task, I chose two Bach chorales. These chorales were chosen because of their similarity in style and difficulty. It is often hard to find two different pieces of music that can objectively said to be similar in level of difficulty. These two chorales, however, were composed using the principles of musical counterpoint. Counterpoint is a set of rules and guidelines that dictate how music can be composed. It is a relationship between the voices of a piece that is interdependent harmonically but independent in contour and rhythm. The two chorales we chose are standardized based on this principle. While they are different pieces of music, they ought to be highly similar in musical difficulty level. They also have an equal number of accidentals (sharps, flats or naturals in the score). Both pieces are the same length, and each has four voices: bass, tenor, alto, and soprano. This allowed for an easy way to segment the music in the adaptive condition in order to progressively increase the difficulty. Due to experimental constraints and skill level of participants, the

pieces were slightly altered. They were transposed into the same key (G Major) and the eighth notes were removed to maintain rhythmic consistency. By removing some of the notes, some minor counterpoint rules are no longer met. However, the underlying structure and form of both pieces is still intact.

Figure 4.1 shows the musical scores for the BACh and normal conditions. In the BACh condition, each level adds a full line of music until it reaches the full piece at Level 4. In the normal condition, learners were told they could learn the piece the way they normally would. Piano players are typically classically taught to play pieces first with their right hand, then their left hand, and finally both together. During that process, it is not uncommon for learners to play one line of the right or left hand, and add further lines as competency grows. As a result, all participants segmented the pieces while they learnt during the normal condition. Most participants started learning with their right hand only, one participant who was very good with his left hand started with his left hand first. Participants then incorporated the other hand as they progressed while learning. This can be viewed as akin to self-judgment of when to increase task complexity while learning.

4.3.1.1 MIDI Keyboard and Bitwig

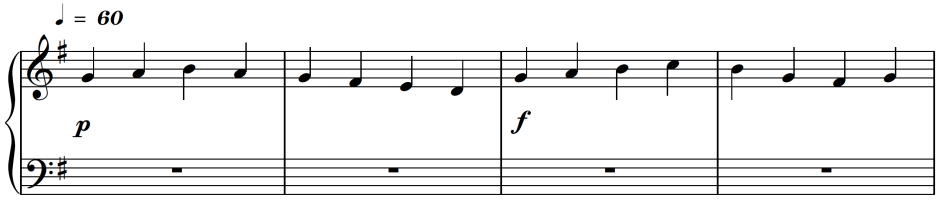
Participants completed tasks on a full-sized Yamaha keyboard with weighted keys. The keyboard was transmitting MIDI data via USB to a computer running Bitwig studio, a digital audio workstation that allows for MIDI recording and playback as well as visual displays of recordings. Participants' musical data was recorded with this equipment for later analysis.

4.4 Algorithms Behind BACh

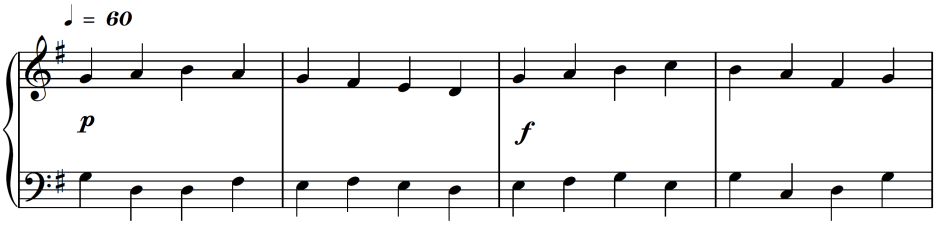
4.4.1 Initial Real Time Classification of Learning

In section 3.1, I discussed the data analysis and classification techniques of the fNIRS system. BACh is the system described in 3.1 *plus* the real-time analysis and adaptations of participant learning based on their cognitive workload.

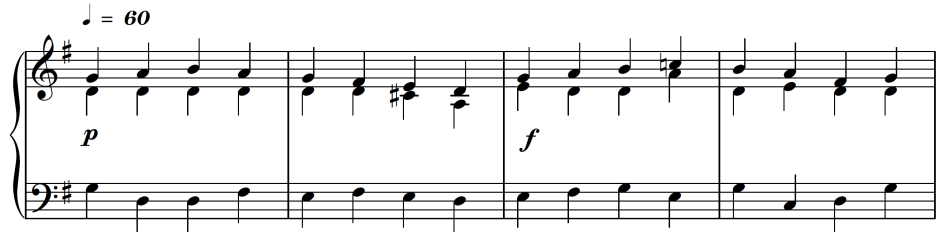
BACh condition - Level 1



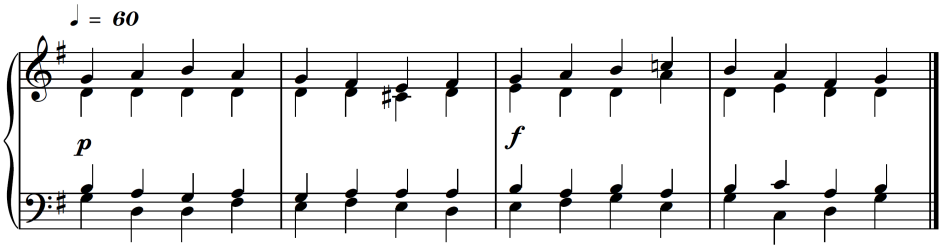
BACh condition - Level 2



BACh condition - Level 3



BACh condition - Level 4



Normal Condition

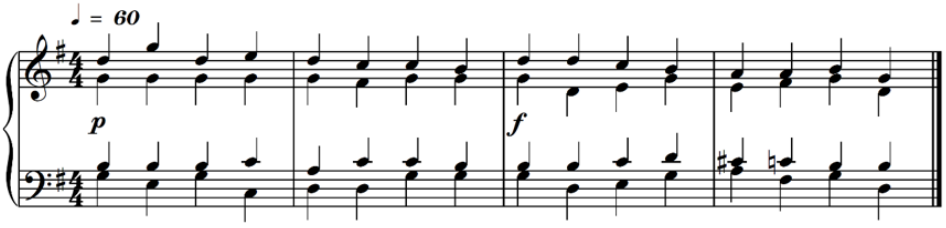


Figure 4.1: Scores used for BACh and Normal conditions. BACh increased the difficulty level by adding a full line of music when users' cognitive workload fell below a threshold, indicating that they could handle more information.

As described in section 3.1, the LIBSVM machine learning model is able to provide probability estimates on how likely the real-time test data belongs to either classification class. Thus, initially, I had BACh analyze the last 30 seconds of real-time fNIRS data in order to calculate a prediction and confidence interval. For each sliding window of 30 seconds, the system provided a prediction and confidence value every 500 milliseconds using the same mean and linear regression slope features described in section 3.1. The overall system confidence averages of cognitive state by averaging the last 8 seconds of predictions. The sliding window interval was based on the trial length from the training tasks.

When the average confidence value for low cognitive workload over the last 10 seconds of predictions was above 80%, BACh would increase the difficulty level of the learning task.

The parameters set for the sliding window interview, x number of seconds of predictions, and the threshold confidence value for an adaptation were initially closely based on real-time studies that had successfully used fNIRS in adaptive systems such as [3].

However, pilot studies demonstrated clearly that this algorithm was not sensitive enough by itself for the varying levels of difficulty presented in BACh. In the next section, I describe how the algorithm behind BACh was developed to reach the final version.

4.4.2 BACh's Algorithm

The pilot studies clearly demonstrated that the algorithm needed to be refined. The switch to the next difficulty level was often occurring either too early or too late. In order to give BACh increased sensitivity for each level, I created two further steps in the algorithm:

1. **Confidence threshold automation**.
2. **Difficulty level specific parameters**.

I now discuss these steps in further detail.

4.4.2.1 Confidence Threshold Automation

In previous adaptive BCIs, the adaptations would occur if confidence averages were above some fixed percentage for low or high cognitive workload levels (e.g. [199, 3]). In this thesis, through a series of iterative design techniques and pilot studies, I implemented flexible confidence thresholds that would respond to each participant individually. In order to do this, I first gathered data on the confidence values of the real-time test data outputted by the LIBSVM machine learning model for a variant length of time determined by pilot studies on each level of difficulty. Once I had a range of data of the confidence values of the participant learning a piece on a specific level, I used a percentile value of from the range collected in the first x seconds. Each percentile was different for each level (discussed further in section 4.4.2.2). BACh then changed difficulty levels if 65% of the last 10 seconds of predictions were above the percentile threshold. This percentage was determined by pilot studies.

By using a percentile and not a fixed percentage I was able to tune BACh more finely to each individual participant. In this way, the thresholds were set at a more accurate and representative level of each user's cognitive workload while learning. This ensured that the system would move onto the next level of difficulty when it was confident that the user could cognitively handle more information.

4.4.2.2 Difficulty Level Specific Parameters

While participants' cognitive workload decreased during each level of difficulty as they mastered each level, their cognitive workload would fall to relatively *different thresholds* on *different levels*. This meant that the same parameters could not be used on different levels as participants' cognitive workload would respond differently at each level. For example, a threshold that worked well for a participant at an easier level, could result in a premature adaptation in a harder level when they need to cognitively work harder for learning to occur. Therefore, each level of difficulty had its own percentile set, with the value falling lower as difficulty levels increased.

Another problem presented by differing difficulty levels that was revealed by

pilot studies was the fact that participants did not always play all of the voices immediately from the start of each level. For example, on level 3 there are 3 lines of music that are to be played concomitantly. However, some participants would not start playing level 3 with all 3 lines at the same time. This resulted in misleading data collection as they were not playing at full cognitive capacity until they reached all 3 lines of music. BACh, therefore, started measuring participants' cognitive workload after they were playing all 3 lines of music. If BACh started measuring their cognitive workload earlier before the full level of difficulty, BACh would quite correctly think that they were ready to move on earlier as it was reading their cognitive workload for an easier difficulty level.

4.5 Evaluation of Dependent Variables

4.5.1 Dependent Variables

I compared the following dependent variables between the BCI and control conditions:

- *Correct notes*: This is the number of correct notes on the musical score given to participants.
- *Incorrect notes*: When an incorrect note is played in the place of a correct note (Figure 4.2). This is a measure of precision or quality of learning.
- *Extra notes*: A note that is a correct repeat or incorrect extra note. This is different from an incorrect note. Often when learners make a mistake they will play the beat again to correct it, this will result in extra notes (see Figure 4.2) and is indicative of incomplete learning.
- *Errors*: A temporal group of notes that is a mistake which includes incorrect or extra notes. This is a more lenient measure as it differs from incorrect or extra notes which are counted individually. For example, a user could play

several incorrect notes in one temporal group, this would be counted as one error (Figure 4.2).

- *Missed notes*: This measure is to account for recall (vs. precision which can be measured by the number of correct and incorrect notes). Some people may play very few notes and prioritize precision (i.e. a high number of correct notes out of total notes) but this would not be an accurate representation of learning. The number of missed notes (Figure 4.2) will account for recall and acts as a measure of completeness.

- *Total time played*: Learners had an unlimited length of time to play the piece once all the way through at the end. The total time it took to play the piece is a good indicator of how well they learned the piece by how fast they could play it.

- *Gap between notes*: This is an indicator of well a piece is learned by how long it takes to move from beat to beat. Incomplete learning often involves hesitation and variance between notes as learners try to move from one group of notes to the next.

- *Beats per minute (bpm)*: A faster tempo indicates increased learning as players can move with more ease from beat to beat. This is different from gap between notes which can illustrate greater variance while bpm shows overall speed. (All pieces were meant to be played at a certain speed which none of the players surpassed.)

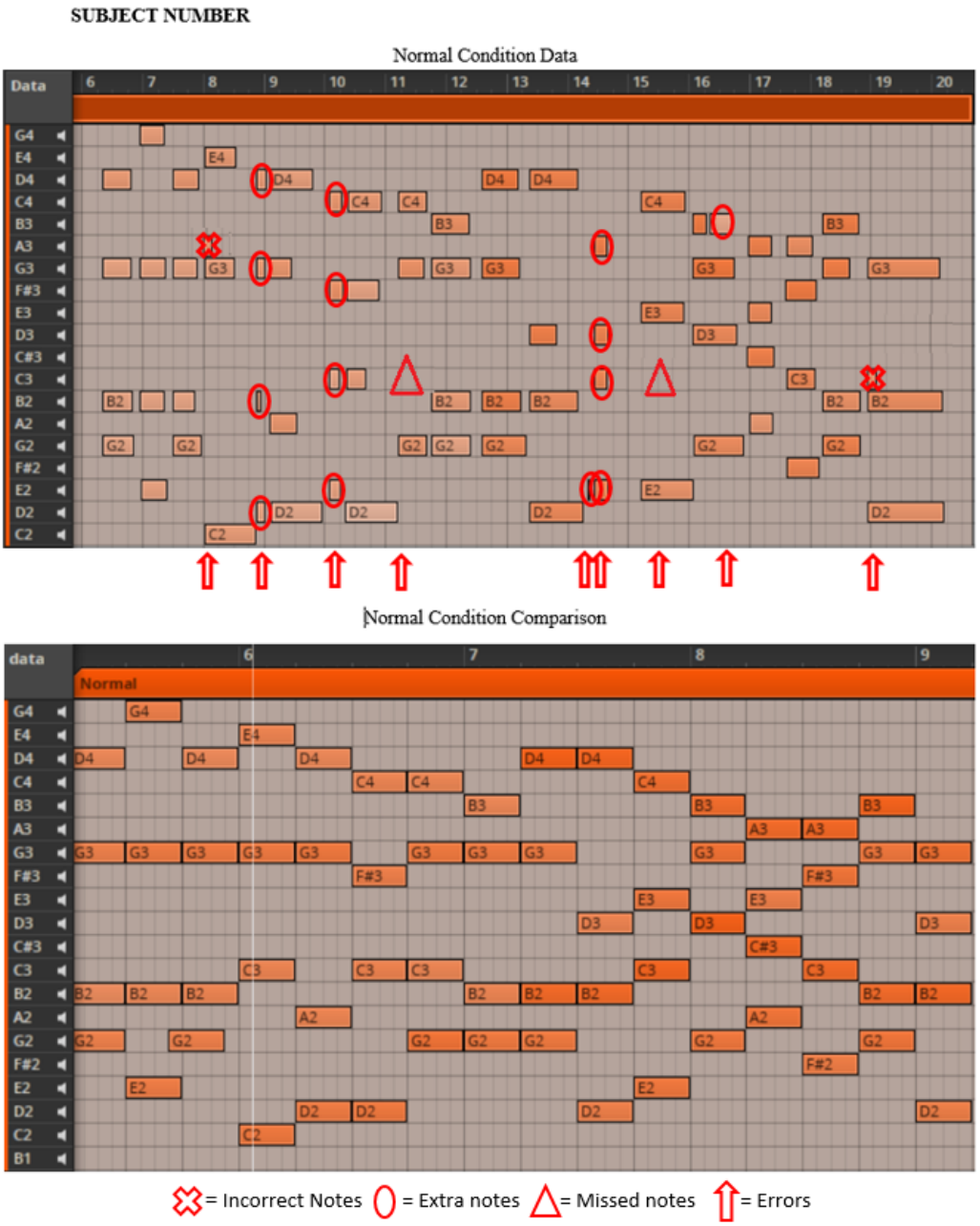


Figure 4.2: Examples of piano rolls from a participant's performance (top) compared with a computer generated groundtruth of the same piece (bottom). Incorrect notes are denoted by a cross, extra notes by a circle, and missed notes by a triangle. An error is a temporal group of notes and is denoted by an arrow pointing up at the column of notes containing an error.

4.5.2 Musical Assessment

The dependent variables total time played, gap between notes and bpm were assessed computationally from the performance data. The others had to be compared to the groundtruth by hand as participants all played at different speeds and score following is an open research problem and beyond the scope of this paper. An expert musician who had taught piano in the past rated participants' learning compared to the groundtruth piano rolls. Figure 4.2 shows piano rolls from an example of a participant's performance compared to a computer generated groundtruth of how the piece should be played. The figure shows how incorrect notes, missed notes, extra notes and errors were identified and counted.

4.6 Experiment 2 - Results

4.6.1 fNIRS Data

We first verify that the brain data demonstrated effective classification of high and low cognitive workload while users played easy and hard pieces on the piano.

Figure 4.3 shows the mean and standard error in the oxygenated hemoglobin (oxy-Hb) of participants while they played easy (blue) versus hard (green) pieces on the piano. We present the mean findings across all participants across all 30 trials in Figure 4.3 to illustrate this general trend.

To investigate differences between hard and easy pieces, we performed a t-test on the mean change in oxygenated hemoglobin. This revealed a significant difference between conditions when participants played an easy piece ($\mu = -0.068, \sigma = 0.077$) versus a hard piece ($\mu = -0.00005, \sigma = 0.124$) on the piano ($t(29) = -2.42, p = .01$). Means and standard errors are shown in Figure 4.3.

The significantly higher levels of oxy-Hb when participants were playing harder pieces on the piano correspond with the hemodynamic literature, whereby, when there is increased cognitive activity in an area of the brain, excess oxygen is provided to that area. The increase in oxygen consumption is less than the volume

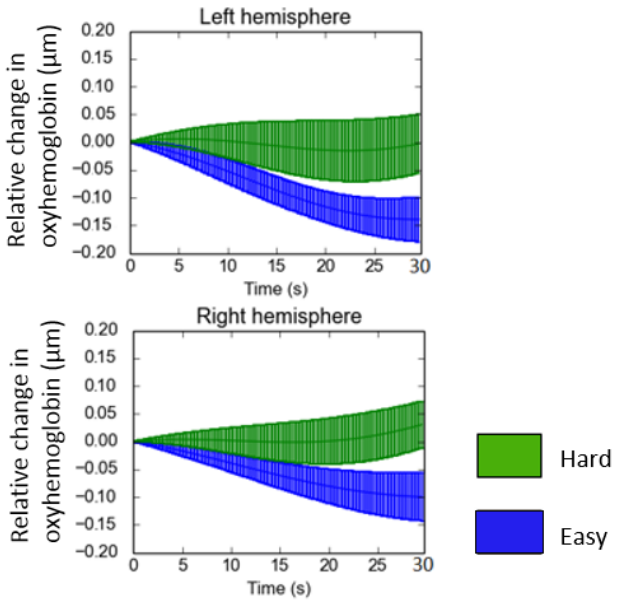


Figure 4.3: Mean change and standard error in oxy-Hb across all participants. Although each participant was modeled individually, the fNIRS signal exhibited a general trend with higher levels of oxy-Hb corresponding with hard pieces. The mean change in oxy-Hb was significantly higher in participants when they played an hard piece than an easy piece ($p = .01$).

of oxy-Hb provided, resulting in more oxy-Hb [55].

4.6.2 Performance Data

I evaluated the performance data of participants using several dependent variables (Table 4.1). We carried out Shapiro-Wilk tests that verifed that the data was non-parametric, followed by Wilcoxon Signed-rank tests on all dependent variables (Table 4.1).

Results indicated a pattern for two traits: 1) *Increased Accuracy*, and 2) *Faster speed* with BACh. We present the results below.

4.6.2.1 Increased Accuracy

Results showed that participants played significantly more correct notes ($Z = -1.9689, p = .05$), missed significantly fewer notes ($Z = 2.3151, p = .019$), played significantly fewer incorrect notes ($Z = 2.4401, p = .015$) and made fewer errors ($Z = 3.0351, p =$

.004) with BACh (Table 4.1). All of these findings are indicative of higher accuracy with BACh.

Dependent Variable	Z	p	effect size
Number of correct notes	-1.9689	**0.05202**	0.304
Number of incorrect notes	2.4401	**0.0153**	0.377
Number of missed notes	2.3151	**0.01911**	0.357
Number of errors	3.0351	**0.003793**	0.468
Number of extra notes	0.8796	0.3633	–
Total time played	2.5337	**0.009186**	0.391
Mean gap between notes	2.482	**0.01099**	0.383
Average BPM	-2.719	**0.00525**	0.419

Table 4.1: Results from Wilcoxon Signed-rank test. Significant results are highlighted in bold. Findings indicate a pattern for increased accuracy and speed in BACh over the normal condition.

The first two graphs in Figure 4.4 represent the number of incorrect notes and number of errors for each participant. The upward sloping lines (blue) are indicative of better performance with BACh with fewer incorrect notes and errors. Downward sloping lines (green) indicate better performance in the normal condition. Horizontal lines that indicate equality are also indicated in green. Only three participants played fewer incorrect notes in the normal condition, and only one participant played fewer errors in the normal condition.

Figure 4.4 reveals that players who are less skilled (i.e. play more incorrect notes or make more errors) tend to benefit more greatly from BACh with steeper slopes demonstrating that there was a larger difference between the two conditions.

4.6.2.2 Faster Speed

Results also showed a propensity towards faster playing speed with BACh. With BACh, participants had a faster overall time in the playing the whole piece ($Z = 2.5337, p = .009$), had a lower time in the mean gap between notes ($Z = 2.482, p = .01$), and had a higher beats per minute (bpm) speed ($Z = -2.719, p = .005$) than in the normal condition.

The last two graphs in Figure 4.4 represent the total time it took to play the

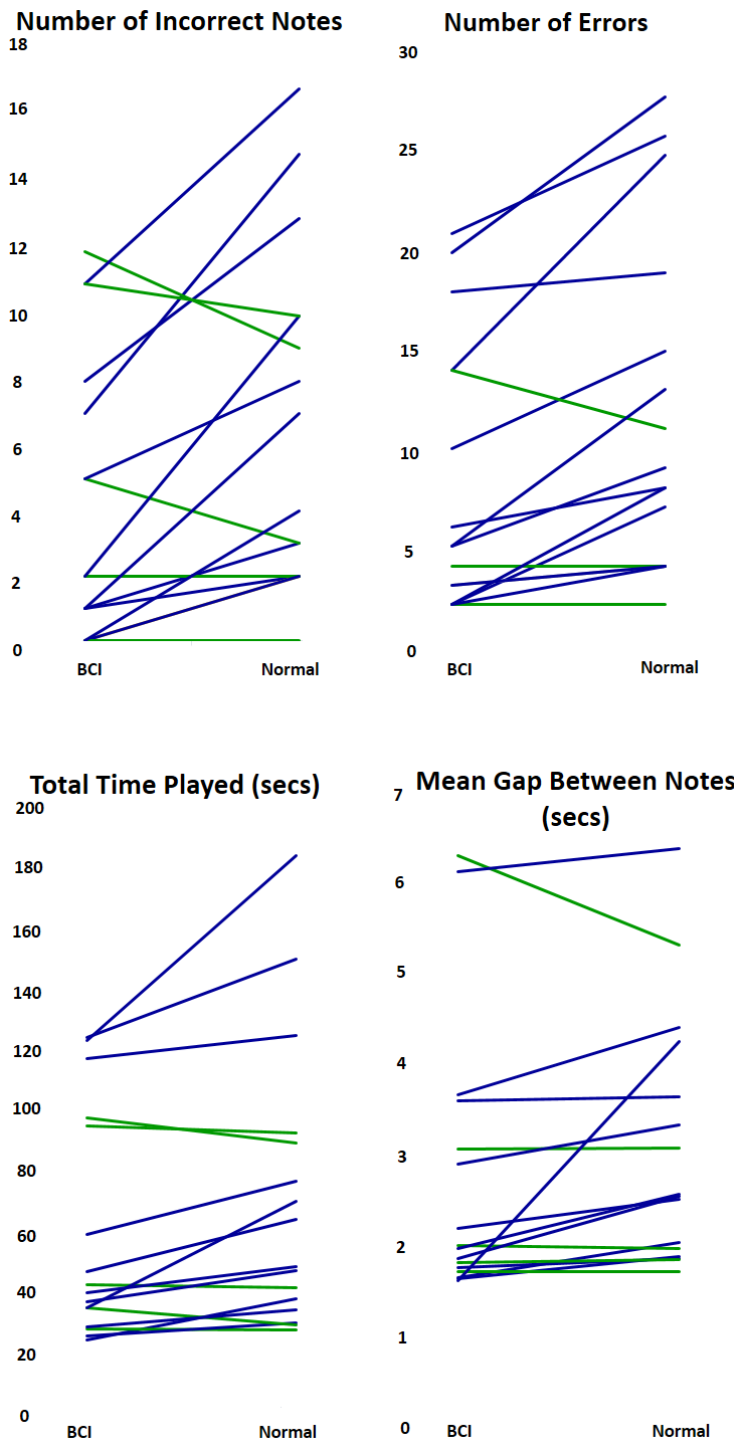


Figure 4.4: Slopegraphs showing the effects of BACh for each participant. Upward sloping lines (blue) are indicative of better performance with BACh. Downward sloping lines (green) indicate better performance in the normal condition. Horizontal lines that indicate equality are also indicated in green. Participants showed significantly better performance with BACh in all four conditions ($p < .01$). Interestingly, participants who were less skilled seemed to benefit more greatly from BACh.

whole piece and the mean time in between notes. The upward sloping lines (blue) are indicative of better performance with BACh with faster overall playing time and less time between notes. Downward sloping lines (green) indicate better performance in the normal condition. Horizontal lines that indicate equality are also indicated in green. Only three participants played a faster time in the normal condition and only one participant had less gaps between notes in the normal condition.

Taken together, these results suggest that BACh helps piano players to play with significantly higher accuracy and speed than the normal condition.

4.6.3 Questionnaire Data

Participants' subjective ratings correspond with their performance. Figure 4.5 shows that participants showed noticeable preferences for BACh in how well they felt they mastered the piece, how correctly they felt they played, and how easy they felt the piece was to learn. The questionnaires can be viewed in Appendix B.

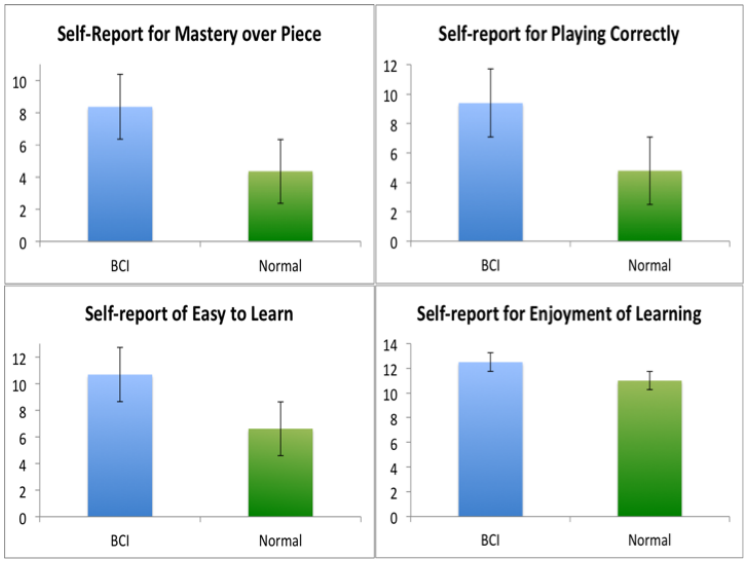


Figure 4.5: Mean and standard error of participants' ratings of BACh and the normal condition. Participants showed noticeable preferences for BACh in how well they felt they mastered the piece, how correctly they felt they played, and how easy they felt the piece was to learn.

4.6.4 Timings of Level Changes

One of the key contributions of this paper is a learning task where difficulty is increased based on participant cognitive workload falling below a certain threshold. I investigated the timing of these changes while users were learning the pieces in further detail in two ways: 1) I interviewed participants and asked them what they thought of the changes in the levels of difficulty, and 2) I investigated the variance in timing data to see if there were individual differences in learning times. I present the results below.

4.6.4.1 Interview Data

After the experiment was over we gave participants a recorded interview where we asked them what they thought of the timing of the changes in difficulty levels. The goal of BACh was to change the level of difficulty at a time when the participants' cognitive workload was low enough to handle more information and a higher level of difficulty. When participants were asked about the timings of these changes, which were controlled entirely by the system, their feedback was generally positive:

> *"I thought it was good timings because by the time I learned, it gave me enough time to learn the individual lines, one by one."*

> *"I thought they were good times for changes, all of them."*

> *"Having a timing system can be jarring; you should only add new things when you know that the person has completed the existing part, but these timings were fine."*

One participant even seemed convinced that the experimenters were triggering the changes:

> *"I wasn't sure if you were controlling it or not because when it was added was a pretty appropriate time for me to add on to a part. Especially because in the beginning, one line, for me at least, is very easy to sight*

read so just getting that melody in my head and figuring out fingering for that one line and then adding on to it very quickly afterwards was helpful. I felt the timing was pretty good. I wasn't sure if it was timed or if you were like, oh she's done with this part, so add on to the second part."

Such comments suggest that BACh was able to effectively support participants in their learning process by increasing difficulty levels when they were ready and able to handle more information. Some participants were not completely satisfied but in cases like these comments were similar to phrases like this:

"Sometimes I wouldn't notice it would change until I would look at the screen; it was a little confusing when I would look up. Yeah, it changed when I had learnt pretty much what I could learn before it changed; it was enough time to learn it."

or

"I thought they [the timings] were pretty good. I think it seemed pretty good overall; the only thing would be the first one was a lot easier to learn because there was only one line, but it wasn't that bad."

The comment about the levels changing without being noticed can be easily fixed with a small notification sound. The comment about the easiest level taking too long can also remedied by increasing the cognitive workload threshold when levels are very easy. Overall, there were no standalone or overtly negative comments about the timings of BACh. Participants seemed to feel the changes were happening at appropriate times, suggesting that we could help manage their cognitive workload as they were learning by altering the difficulty of a given stimulus.

4.6.4.2 Individual Differences in Level Changes

While participants spent less time on easier levels and longer lengths of time on levels with increasing difficulty, we did find individual differences in time spent *within*

levels. The length of time (in seconds) spent on level 1 ($\mu = 58.58, \sigma = 33.47$), level 2 ($\mu = 185.60, \sigma = 64.62$), level 3 ($\mu = 248.49, \sigma = 151.59$), and level 4 ($\mu = 407.32, \sigma = 179.80$) had high standard deviations, suggesting that BACh was responding to individuals' cognitive workload and learning abilities rather than dishing out a standard formula of learning time for everybody.

Figure 4.6 shows the variation in time within levels for each participant and the general pattern of longer time spent on more difficult levels.

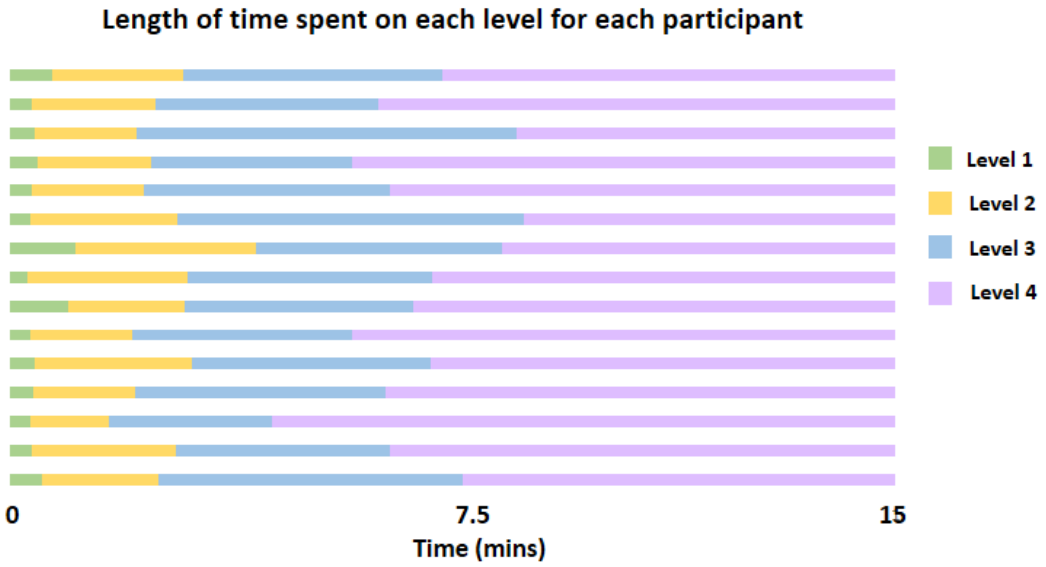


Figure 4.6: Length of time spent participants spent on each level of difficulty. While participants spent increasing lengths of time as level difficulty increased, there were individual variation within levels, suggesting that the system was responding to user cognitive workload for each participant individually.

4.7 Experiment 2 - Discussion

We presented an adaptive brain-computer interface, BACh, that increases difficulty in a musical learning task when learners' cognitive workload fell below a certain threshold. We showed that BACh significantly increased speed and accuracy compared to a control condition where participants learned pieces the way they normally would. Participant feedback also demonstrated that they felt that they played bet-

ter and it was easier to learn with BACh. Participants felt that the timings of the difficulty level changes were accurate. Results showed that while there was a general pattern of participants spending longer time on the more difficult levels, there was individual variation on time spent *within* levels, indicating that BACh was responding to individual differences.

I now discuss the challenges of modeling and adapting to learner cognitive workload using brain sensing and designing adaptations accordingly, the importance of responding to learners individually, and designing learning systems based on learner expertise.

4.7.1 Modeling and Adapting to Cognitive Workload

When designing BACh, I was extremely careful to first do no harm to the learning process. I did not want the adaptations made by BACh to impede or interrupt periods of learning in the user. I implemented this by avoiding disruptive adaptations during periods of cognitive workload that indicated possible learning.

One of the limitations of physiological computing is that of mapping from physiological signals to psychological states [51]. For example, if the fNIRS data signals that the learner's cognitive workload is high, this could be indicative of either a phase of learning where the user is being pushed cognitively in a constructive way, or it could indicate that the user is overloaded or overwhelmed and is not able to learn in that state.

I purposefully designed BACh not to interfere when learners' cognitive workload was high in order to avoid disrupting a period of possible learning. This is the reason why the adaptations never *decreased* the difficulty level during periods of high cognitive workload, as this could have proven both disruptive and frustrating to the learners. However, using high cognitive workload as the determining factor in adaptive learning systems would be a very interesting topic to investigate further. The discussion on whether high cognitive workload is 'good' or 'bad' during learning is still an open research problem. It may be that the solution lies in identifying learner emotion, or affect, in conjunction with cognitive workload, to reveal a higher

dimensional mapping of the learner's state.

By examining the performance data and participant feedback, it seems that BACh was able to correctly respond to learner cognitive workload to improve playing speed and accuracy. This suggests that BACh was able to accurately model and adapt to learner cognitive workload in a learning task and guide learners into the zone of proximal development.

4.7.2 Responding to Learners Individually

BACh responds to each learner individually. In the early stages of development, BACh was originally designed to increase difficulty levels during learning when learners' cognitive workload fell below a *fixed percentage threshold*. Early studies showed us very quickly that this was not going to work. Each learner was very individual in their brain measurements, and what worked very well for some, would not even trigger a fixed percentage threshold for others. Furthermore, even if it worked well for one level, it certainly was not guaranteed that it would work for levels of varying difficulty.

Through a series of iterative design techniques, extensive participant feedback, and many pilot studies, I came to an algorithm that would assess learner cognitive workload using both the learner's brain data from the training task and brain data from the *current* level of difficulty that the learner was on to create a threshold for that individual on that level of difficulty.

While there is room for further improvement, BACh's algorithm significantly increased learner speed and accuracy compared to how learners would normally learn a piece of music. Some of this individual responsiveness of BACh can be seen in the variation of the length of time taken by learners on each level. While there was a general pattern of longer periods of time spent on levels with increasing difficulty, there were still individual differences *within* each level. Individual differences in learning is a significant and complex topic and it is important for any CBE system to take it into account.

4.7.3 Expertise of Learners

Learning and CLT literature has explored a phenomenon called *the expertise reversal effect* whereby instructional techniques that are beneficial to beginners can have the reverse effects on more experienced learners [88, 89]. This difference between expert and novice is thought to be due to experts' previously acquired schemata that have built complex elements into fewer elements. In a task, the novice will have to use more of their limited working memory to access and process the individual elements than the expert, reflecting in the difference in skill.

The results actually suggested that less skilled piano players benefited more from learning with BACh, as demonstrated by the steeper lines in Figure 4.4 for players who made more mistakes. BACh was designed for beginner piano players. We had originally recruited both beginner and intermediate piano players during our pilot studies, but quickly found that both the musical pieces and the algorithmic parameters of BACh were too easy and frustrating for intermediate players.

I foresee a similar system being of use to intermediate players with harder pieces to learn and different parameter settings for threshold changes. We can also see such a system being frustrating to even more advanced players who might prefer to learn their own way and have access to the whole piece from the beginning. I, therefore, simply suggest that BACh is targeted at and useful for *beginner* piano players.

4.8 Experiment 2 - Conclusion

I presented a brain-computer interface, BACh, that measured cognitive workload using brain sensing and increased difficulty levels when learners' cognitive state indicated that they could handle more information. I showed that participants learned musical pieces with increased accuracy and could play with faster speed with BACh compared to a control where they learned the way they normally would. Participants also commented that they felt that they played better and they liked the timings of the changes. We also found individual variations in the time spent

within each level of difficulty, suggesting that it is important for tutoring systems to respond to individual differences and needs.

I designed BACh for beginner piano players, however, the underlying premise of BACh can be applied to any learning situation for a complex task that can be broken down into steps of increasing difficulty. I address the generalizability of BACh in section 7.2.1. By using brain sensing, BACh is able to acquire an objective view into the learner's cognitive state and adjust the learning task to best guide the learner into the zone of proximal development.

Chapter 5

Experiment 3: Creativity

This work was published at NIME 2015. Yuksel et al. (2015) BrAAHMS: A Novel Adaptive Musical Interface Based on Users Cognitive State. *New Interfaces for Musical Expression (NIME).* pp 136–139 [230].

5.1 Experiment 3 - Introduction

Brain-computer interfaces (BCIs) have been used to increase the effective communication bandwidth between the human and the computer by passively obtaining and using extra information about the user [3, 199]. In a musical instrument interface, such as a piano keyboard, the user's communication channels may be limited by the paucity of input devices, body parts, and attention. An adaptive BCI can increase this bandwidth by providing a cognitive communication channel that passively measures user cognitive state without requiring additional effort or attention on the part of the user. This can then be used as an additional input channel that implicitly modifies the music at appropriate times.

Traditionally, most musical BCIs have been based either on different variations of mapping brainwaves to soundwaves [130, 138, 140] or by using brain signals explicitly to modify aspects of the music [27, 72, 211]. For example, the magnitude spectrum of the brain signals are used to shape the spectrum of musical frequency [130, 138, 152] or certain features are mapped to certain musical features [140, 226].

More recently, the power spectrum of EEG signals have been used to map algorithm selection to create music [140]. However we seek the ability to go beyond the lower level signal and to extract meaningful information from the brain data about cognitive activity [138, 141, 180, 191].

Recent events in the field of brain sensing have opened up a new avenue of measuring and using higher level meaningful cognitive data implicitly without any extra attention required from the user. Cognitive workload, which has been measured by fMRI [36], EEG [64] as well as fNIRS [12], has led to BCIs that can implicitly adapt to user cognitive state in real-time [3, 199]. Solovey et al. [199] demonstrated the first real-time BCI that adapted to users' cognitive state with fNIRS. Ayaz et al. [12] and Afergan et al. [3] used cognitive workload to assist user load in unmanned aerial vehicle (UAV) simulations with fNIRS.

BCI technology has now reached the point where this is an opportune time to introduce a musical system that implicitly adapts in real-time to users' cognitive state. I introduce **BrAAHMS: Br**ain **A**utomated **A**daptive **H**armonies in a **M**usical **S**ystem. BrAAHMS is a real-time musical BCI system that calculates cognitive workload to implicitly adapt to users in the creative task of musical improvisation. BrAAHMS adapts to the user's cognitive workload by adding or removing musical harmonies that are related to the notes that the user is currently playing, hence augmenting their music without altering their original general direction.

In section 2.3.1, I discussed how there has been a clear need for creativity support tools and creativity research, and how brain sensing has been highlighted as a technology that can provide information that other methodologies, including other physiological ones, cannot [122, 191, 192].

However, the challenge that has been identified is that it is not yet clear *how* a system that measures brain state could aid user creativity. This chapter addresses that problem by using a brain-computer interface to detect user cognitive workload and adapt to participants improvising on the piano in real-time to aid creativity.

5.2 BrAAHMS: Iterative Designs

Experiment 1 established that there were differences in the fNIRS signal when participants played easy vs. hard pieces on the piano. I therefore built upon these findings to design and create a novel musical brain-computer interface BrAAHMS: BRain Automated Adaptive Harmonies in a Musical System.

5.2.1 Design Iterations and Pilot Studies

Prior to designing the musical adaptation used in Experiment 3, a number of design iterations and pilot studies took place. These iterations explored the space of possible musical additions that an adaptive system could provide, and resulted in a final design that selected the musical enhancements indicated as most helpful or pleasing by users in the pilot studies.

5.2.1.1 Preliminary Interviews

I discussed the topic of a brain-based adaptive musical interface with one professor of music and engineering, one professor of music and machine learning, and several music graduate students before designing and building the system. The main outcomes were that 1) expert musicians would not appreciate interference in their creativity, and 2) most people who take piano lessons are classically trained to play musical scores, but are not generally taught how to improvise (this was also later reflected in participants' improvisational experience in Experiment 3). I therefore decided to focus on building a system designed to aid beginner to intermediate pianists at musical improvisation.

5.2.1.2 Melodic fragments from all piano pitches

The starting point of the iterative design process was a system that inserted precomposed melodic fragments while the user was playing, making use of any pitches present on a standard piano keyboard. I determined that this form of musical addition did not complement the user input and was instead experienced as contrarian,

such as the following report from a participant: "*[The additions] made composing/improvising much more difficult because it distracted my hearing and what I was playing and distracted me*". This starting point had two main drawbacks. First, since the melodic fragments were precomposed, they did not tend to coincide with the users' playing. Second, the unrestricted availability of all pitches on the piano resulted in pitch additions that altered the harmonic content of the users' music in a confusing way.

5.2.1.3 Layered melodic fragments with harmonic consistency

To address the confusing harmonic alterations, a restriction was added to force all system additions to match the same harmonic center (or "musical key") as the user input. For example, in the key of C major, only white keys on the piano are available to the system. To further refine the exploration, four layers of musical additions were created and tested. The layers were:

- Simple melodic additions in conjunction with the user input, which used a subset of pitch classes from the major scale (e.g. for a piece in the key of C, using a subset of the white keys only).
- More complex melodic additions, which played more notes faster in conjunction with the user input.
- Harmonic additions to the melodic portion of the user input, using harmony rules derived from Western music theory.
- Harmonic bass additions to accompaniment portion of the user input, using harmony rules derived from Western music theory.

As the user maintained a low level of measured cognitive workload, additional layers were added every ten seconds, starting with simple melodic additions and ending with all four layers combined. If the measured cognitive workload became higher, layers were removed.

User feedback on these musical additions was more positive than the precomposed fragments from earlier. For example, one user wrote: "*I really enjoyed playing the simple melody when each of the other melodies/harmonies came in - it led to*

some really pretty sounding music that I felt I was controlling with only a few keys!". However, other users still found the melodic portions intrusive, specifically with respect to their experience of the tempo and rhythm of the composition. Another participant wrote: "*Sometimes the timing of the additions seemed to clash with the speed I was trying to play*". Overall, the aggregated feedback suggested that the melodic additions were experienced as mixed positive and negative by participants, whereas the harmonic bass additions were consistently experienced as pleasing enhancements to the music created by the user without compromising creative control.

This iterative design process made it clear that the best approach for the adaptive system was to use a simple harmonic addition to reinforce the user input while making sure to respect the rhythmic and tempo choices of the user. This formed the basis for BrAAHMS.

5.3 Final Design for BrAAHMS

5.3.1 Musical Real-Time Adaptation

The design iterations and pilot studies made it clear that the most pleasing adaptation is to harmonically supplement the music created by the user while staying true to the melodic and rhythmic choices provided by the user. The musical adaptation system accomplishes this with a simple harmonic addition that re-emphasizes the melodic pitches and reinforces the harmonic center provided by user-selected pitches, matching the user note onset times to maintain the same rhythm as the user. The right-hand user input (higher pitch middle C or above) typically plays the more melodic portion of a piece. For the right-hand input, the system adds the pitch one octave above the user pitch (+7 notes of the major scale). The left-hand user input (lower pitch below middle C) typically plays the accompaniment or bass part in a piece. For the left-hand input, the system adds two pitches: 1) one octave below the user input pitch (-7 notes of the major scale), and 2) the third of the chord between the octave and the user pitch (-5 notes of the major scale). These intervals are shown in Figure 5.1. These musical additions are consistent with typical

Western harmonies [169].

The octave additions serve to re-emphasize the pitches chosen by the user. The addition of the second left-hand note provides a harmonic reinforcement of the tonal center specified by the left-hand user input. The added pitch is the third of the chord (e.g. if the user plays a C, then the added note will be an E). [169]. Adding this enriches the harmonic texture of the musical sound. The third of the chord is specific to the harmonic mode of the music overall (e.g. if the user plays a C, the added E implies C major rather than C minor), so adding the third of the chord can help maintain a harmonic consistency across different user inputs. This adds musical depth to the harmonic content chosen by the user.

By providing these simple harmonic additions in this way, the system adds depth and richness to enhance the musical choices made by the user. The user retains a great deal of flexibility to make melodic choices with their right-hand, and retains complete control of the rhythmic aspects of the music creation experience.

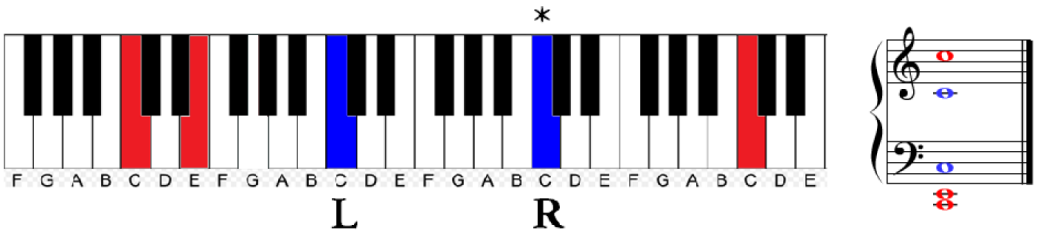


Figure 5.1: The musical harmonies added by BrAAHMS. a) Blue notes denote the keys the user is pressing. Red notes denote the additions made by the system. The L and R depict the fingers of the left and right hand respectively. The * indicates middle C. b) Music notation showing a sample user input (blue) with the additional pitches provided by the system (red).

5.3.2 Sound Engine

Figure 5.2 illustrates the sound engine I built behind BrAAHMS. The musical additions were implemented in Pure Data (PD) with a piano soundfont through the software synthesizer FluidSynth. QjackCtl was used to connect all systems on a

Linux Operating System. Subjects played their chosen notes on an electronic piano keyboard, which sent MIDI inputs to PD. The combined harmonic MIDI outputs along with users' original notes were generated in PD and sent to the keyboard via the soundfont on FluidSynth which converted the MIDI outputs into audible piano sounds. PD only added or removed musical additions when it received a marker from the fNIRS system, the real-time classifications of which are discussed below.

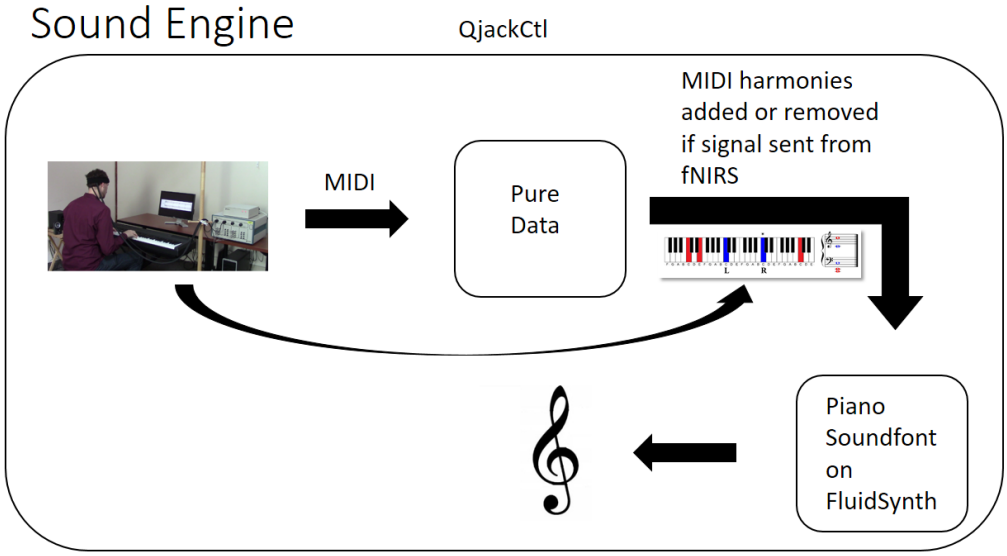


Figure 5.2: Sound engine behind BrAAHMS. Musical additions are implemented in Pure Data through a piano soundfont on FluidSynth. Musical additions were only added or removed when a marker was received from the fNIRS system.

5.3.3 Musical Guidance for Participants

As it is challenging even for expert musicians to improvise or perform in a completely open-ended musical context, participants were provided with a simple piece of notated music to use as a basis for their musical creation. Participants were invited to improvise beyond the confines of the simple skeleton provided by the notated music. The notated music is shown in Figure 5.3. The notation outlines a repeating I-IV-V-I musical chord progression, which is one of most prevalent chord progressions in Western music and is likely familiar and aurally pleasing to participants [76]. In the

key of C, the chords are: C-F-G-C. This musical skeleton provides an easy way for participants to approach the musical creation task.

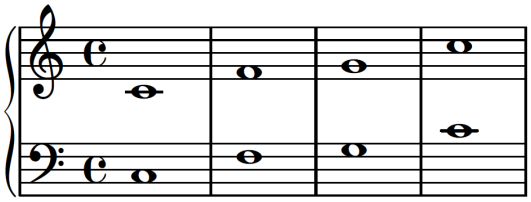


Figure 5.3: Notated music provided as guidance only to participants during musical improvisation.

5.4 Algorithm Behind BrAAHMS

5.4.1 Initial Real-Time Classification

BrAAHMS uses the initial data analysis and classification techniques described in detail in section 3.2 *plus* real-time analysis and adaptations based on user cognitive workload.

As described in section 3.1, the LIBSVM machine learning model is able to provide probability estimates on how likely the real-time test data belongs to either classification class. Thus, initially, I had BACh analyze the last 30 seconds of real-time fNIRS data in order to calculate a prediction and confidence interval. For each sliding window of 30 seconds, the system provided a prediction and confidence value every 500 milliseconds using the same mean and linear regression slope features described in section 3.1. The overall system confidence averages of cognitive state by averaging the last 8 seconds of predictions. The sliding window interval was based on the trial length from the training tasks.

When the average confidence value for low cognitive workload over the last 10 seconds of predictions was above 80%, BACh would increase the difficulty level of the learning task.

The parameters set for the sliding window interview, x number of seconds of predictions, and the threshold confidence value for an adaptation were initially

closely based on real-time studies that had successfully used fNIRS in adaptive systems such as [3].

However, similar to BACh's algorithm, pilot studies demonstrated clearly that this algorithm was not sensitive enough by itself for the varying levels of cognitive workload presented by different individuals during musical improvisation. In the next section, I describe how the algorithm behind BrAAHMS was developed to reach the final version.

5.4.2 BrAAHMS' Algorithm

One of the findings from the pilot studies while designing BrAAHMS was that a fixed threshold for confidence average values did not work for all individuals during musical improvisation. In previous adaptive BCIs, the adaptations would occur if confidence averages were above some fixed percentage for low or high cognitive workload levels [199, 3]. In this work, I automated this threshold value for each individual by setting the algorithm to measure the range of confidence values presented in the first non-adaptive condition during musical improvisation and subsequently set the threshold at the 75th percentile of confidence values for both high and low cognitive workload classifications. In this way, the thresholds were set at a more accurate and representative level of each individual's cognitive workload while improvising. This ensured that the system would only add or remove musical additions when it was confident in the user's cognitive state.

5.5 Evaluation of BrAAHMS

I evaluated BrAAHMS in Experiment 3 over 4 different conditions on 20 participants. I investigated the technical aspects of BrAAHMS as well as user preference through subjecting rankings and post-experiment interviews .

5.5.1 Experimental Design

In order to evaluate user preference, we tested BrAAHMS over 4 different conditions of musical adaptation. The adapatations were to add and remove musical harmonies as discussed in section 5.3.1. However, we did not assume whether to add and remove the harmonies when cognitive workload was high or low to account for individual differences and to investigate user preference. We therefore tested 2 different BCI conditions as well as 2 non-BCI conditions:

- *BCIlow*: Musical harmonies are added when brain signals correspond with low cognitive workload and are removed when brain signals correspond with high cognitive workload.
- *BCIhigh*: Musical harmonies are added when brain signals correspond with high cognitive workload and are removed when brain signals correspond with low cognitive workload.
- *Constant*: Musical harmonies are always present.
- *Non-adaptive*: There are no musical harmonies.

Users first carried out the training task described in Experiment 1 in order to build a machine learning model from their brain data. They then carried out the 4 conditions described above in random order except for the non-adaptive condition which had to come first to extract the automated threshold for low and high cognitive workload. Participants had no knowledge that some conditions were brain-controlled or not as they wore the fNIRS sensors throughout the experiment and were simply asked to improvise and that this would be repeated 4 times. Each of the 4 conditions were filmed with the participants' consent.

I also carried out and recorded post-experiment interviews during which participants watched each condition back on video and were asked three questions as they watched each condition: 1) *Do you have any comments about this trial?* 2) *Do you have any comments about the additions in this trial?* 3) *Do you have any other comments?* They were still blind to differences in conditions. At the end of the

interview they filled in a short questionnaire where circled their favorite trial out of the 4 options: 'System 1' 'System 2' 'System 3' and 'System 4' and then rank each trial with 1 being their favorite and 4 being their least favorite.

5.5.2 Participants

Twenty participants took part in the second experiment (14 female, mean age of 21, SD of 1.9) and were compensated $20. I requested beginner to intermediate pianists. Mean length of time participants had played piano was 7 years (SD 5 years) while 10 of them no longer played piano and 4 only played once a year. Seven had never improvised and seven only improvised rarely.

5.5.3 Experiment 3 - Results and Discussion

5.5.3.1 Technical evaluation

The fNIRS data in Experiment 3 was consistent with the findings from Experiment 1 whereby brain signals correlated with high vs. low levels of difficulty when participants played the piano (Figure 5.4). Similar to our findings in Experiment 1, we performed a t-test on the mean change in oxygenated hemoglobin which revealed a significant difference between when participants were playing an easy piece ($\mu = -0.1, \sigma = 0.1$) versus a hard piece ($\mu = 0.1, \sigma = 0.1$) on the piano ($t(18) = -4.50, p < .001$). The means and standard errors of these conditions are shown in Figure 5.4.

The increase in oxy-Hb during the hard pieces and decrease in oxy-Hb during the easy pieces are once again consistent with harder tasks resulting in higher levels of oxy-Hb [55].

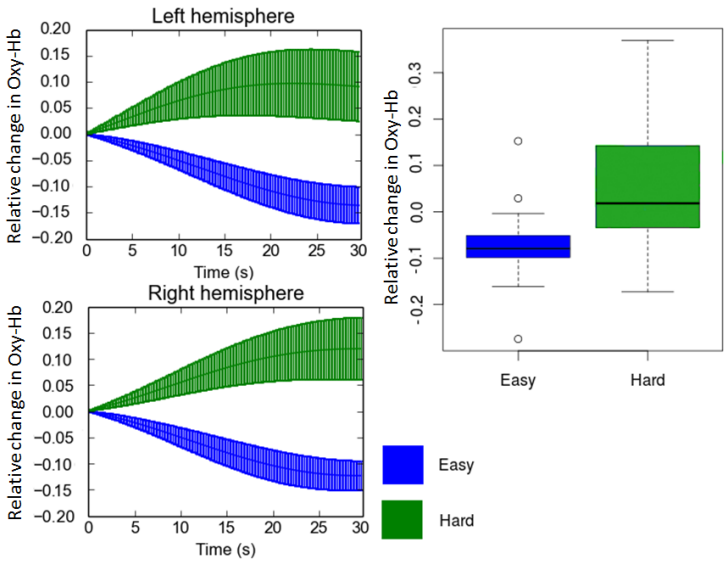


Figure 5.4: Left: Mean change in oxy-Hb and standard error across all trials and participants. Although each participant was modeled individually, the fNIRS signal exhibited a general trend with higher levels of oxy-Hb corresponding with hard pieces, consistent with the findings of Experiment 1 (Figure 3.10). Right: mean change in oxy-Hb was significantly higher in participants when they played an hard piece on the piano versus an easy piece consistent with the findings of Experiment 1 ($p < .001$, also see Figure 3.10).

5.5.3.2 Participants Preferred the BCI conditions

Participants were asked to state their favorite trial at the end of the post-experiment interview, after they had watched footage of each trial. They were blind to the conditions of the trials and wore the brain sensors during all trials.

Fifteen out of twenty participants responded that their favorite condition was a brain-controlled condition. Figure 5.5 shows the favorite conditions of participants.

To investigate why participants preferred the BCI conditions, we turn to the interview data. Both BCI conditions were preferred by different individuals, and while we discuss this in terms of musical expertise, the discussion of why the BCI conditions were preferred in general over non-BCI conditions can also help shed some light on this.

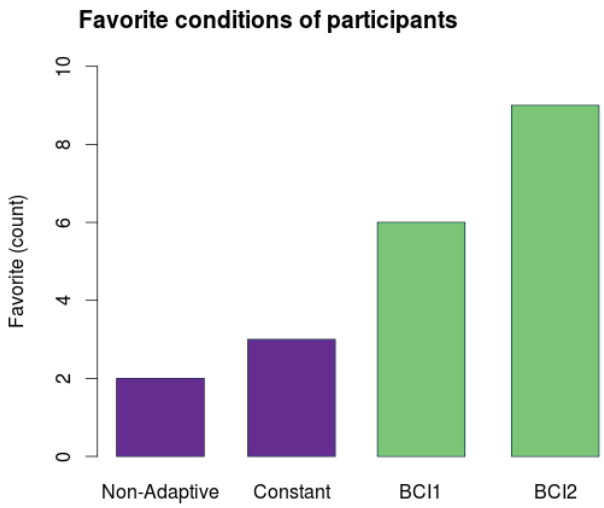


Figure 5.5: Favorite conditions of 20 participants

5.5.3.3 Participants felt more creative in their favorite BCI condition

Findings from interview data revealed that out of the 15 participants who ranked a BCI condition as their favorite, 12 of them commented that the BCI additions helped them to musically perform in a way that they would not have been able to do by themselves. In contrast, only 2 participants made such comments about the constant condition, 1 of which had ranked the constant condition as their favorite. Participants commented that they felt more creative during their favorite BCI condition, such as:

"*[BCIhigh] was my favorite. I felt I was at my creative peak. Obviously it's clunky because I'm not a real pianist, but I felt like I was playing a real song. I was combining my comfort zone with the additions.*"

"*Being that I am not as experienced and I'm able to get a chord out of just one key, I can do sounds that I probably wouldn't know how to make on my own.*

In addition, we present some quotes on how subjects' favorite BCI condition would help them creatively in contrast with their constant condition. One subject found the additions helpful in their favorite BCI condition:

"*During [BCIhigh] I thought the additions were pretty helpful. It's hard to say how. Maybe because introducing another note leads you in a slightly different direction*

than you were heading originally, that can be nice."

However the same subject did not like the constant condition:

"*I thought that sometimes the additions were confusing because I thought that the note that I was trying to play wasn't the note that was coming out. For some reason in this last trial I started to notice them [the additions] more.*"

Similarly, another subject commented that their favorite BCI condition helped them develop new ideas:

"*I liked the additions, I felt like they almost made it easier to come up with ideas because I was staying in the same octaves before and I was like 'Oh, it sounds pretty when its outside of that'.*" However the same subject did not feel the same way about the constant condition:

"*I felt like there were a lot of them [additions]. I remember at least in the beginning it felt like every note, it was kind of weird to have every single note [with additions].*

These findings suggest that the favorite BCI conditions helped participants achieve a musical level that they would not achieve by themselves or be taken in new directions that they would not have alone. The high number of participants (12 out of 15) who all made similar unprompted comments on their favorite BCI condition in response to open-ended interview questions suggests that this was an important factor in their preference. The contrast in comments between the favorite BCI condition and the constant condition within subjects suggests that adaptive musical additions in response to cognitive state could take participants to a creative place that simple, constant musical additions could not.

5.5.3.4 Participants felt their favorite BCI condition was responsive

Out of the remaining 15 participants who stated a BCI condition as their favorite, the other reason that was given for preference was the responsiveness of the musical additions. Comments on their favorite BCI condition included:

"*[BCIhigh] was my favorite... I felt like I was responding to the additions and they were responding to me*"

"*[BCIlow] was my favorite of all of them. I think because I didn't know if*

the additions were going to come in and out it was fun and felt very fluid with the combination of just regular piano [and additions]. I felt that in this one it was responsive."

However in contrast the same subject felt that the their constant condition was not responsive:

"*In [constant] I remember that the additions took a bit more adjustment, I didn't feel that they were as responsive because, I would be in the middle of a very quiet part and all of a sudden it would be doubled and it would be jarring.*"

Another example from a subject who felt that their favorite BCI condition was responsive to them:

"*In [BCIhigh] for some reason I felt that the changes were more responsive and I couldn't tell you why... I couldn't quite figure out any pattern of when they were coming and out so I couldn't tell if I was influencing them but for some reason it didn't feel random, I dont know why.*"

And this same subject felt that the other (non-favorite) BCI condition was not responsive:

"*The additions generally seemed random coming in and out... they didn't seem responsive at all.*"

These comments suggest that responsiveness is an important factor in participants' preference of a condition. This is supported by the fact that there were several comments made about other conditions that were *not* the favorite where responsiveness was lacking. It seems therefore that it is not enough for musical additions to be arbitrary, that they must subjectively feel *responsive* to a user, even if the user does not understand why.

5.5.3.5 Years of experience affected preference of conditions

Of 20 participants, 5 did not rank a BCI condition as their favorite, instead picking either the Constant (3 participants) or Non-Adaptive (2 participants) conditions as their preferred adaptation strategy. We noticed an interesting characteristic in this group in that 4 of these 5 participants had more years of experience playing the

piano (17, 13, 12, and 10 years) compared to the median of 6.5 years of the total group. To investigate these results statistically, we divided the participants into two groups based on years of experience via a median split (at 6.5 years) and examined their rankings of the conditions.

Analyzing the participant's rankings using a Friedman test as the data was nonparametric. There was a statistically significant difference in both the low experience $\chi^2 = 25.74, p < 0.0001$) and high experience groups ($\chi^2 = 27.96, p < 0.0001$). To investigate relationships between conditions, we performed post hoc comparisons using the Wilcoxon rank-sum tests (due to the nonparametric nature of the data), and corrected for the multiple comparisons using the Bonferroni technique, resulting in a significance level set at $p < 0.01$ [70]. Figure 5.6 shows the resulting rank-sums and significant differences.

For the participants with low ($\leq$ 6.5 years) experience ($n = 10$), post hoc comparisons show that participants significantly preferred both BCI conditions to the Constant and Non-Adaptive conditions ($p < 0.001$). There was no significant difference between the BCIlow and BCIhigh conditions ($p = 0.18$), nor was there a difference between the Constant and Non-Adaptive conditions ($p = 0.59$).

In contrast, for participants with high ($>$ 6.5 years) experience ($n = 10$), post hoc comparisons revealed that participants significantly preferred the [BCIlow] condition to both the Non-Adaptive and Constant conditions ($p < 0.001$). However, the [BCIhigh] condition was not significantly different than either the [BCIlow] ($p = 0.04$) (due to the conservative threshold of the Bonferroni correction) nor the Constant ($p = 0.31$) conditions.

The results of these analyses suggest that participants with more experience of playing the piano show more variability in their preferences for adaptations. In contrast, participants with less experience playing the piano show a clear preference of the BCI conditions over the constant and non-adaptive conditions. This is most likely due to more experienced pianists' ease with musical improvisation due to their greater facility with the unadorned piano. It does suggest that such a system would benefit less experienced users and perhaps would rather interfere more with

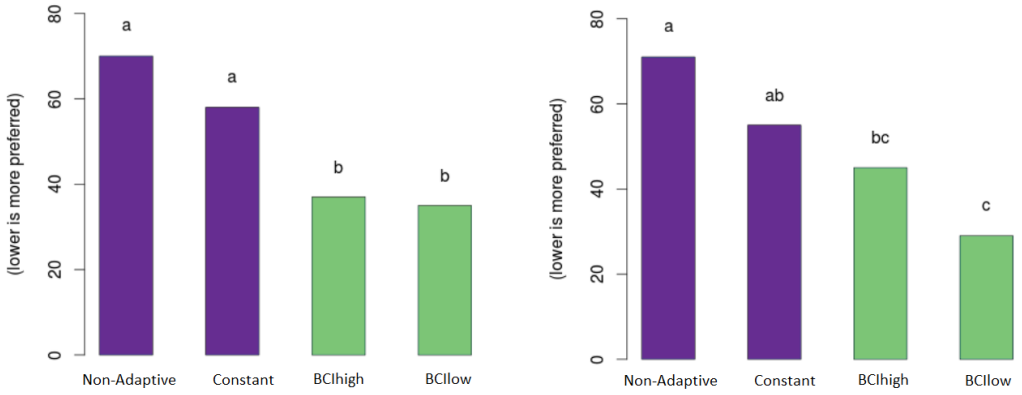


Figure 5.6: Rank-Sums of the conditions based on a median split (6.5 years) on years of experience playing the piano. Rankings of participants with lower experience (< 6.5 years, $n = 10$) (left) and and higher experience (> 6.5 years, $n = 10$). Lower bars indicate a more preferred condition, and bars with the same letter are not significantly different.

more experienced players. This is consistent with our original target audience for BrAAHMS, which had been always been less experienced players.

5.6 Experiment 3 - Discussion

We believe this is one of the earliest demonstrations and evaluations of a real-time brain-based musical system that adapts to brain signals that correspond with cognitive workload. This need for semantically meaningful brain data in respect to musical and/or creative tasks has been highlighted as a need by experts in both the fields of Music and BCI [180, 141, 138] and HCI and Creativity[122, 191].

The fact that most users ranked the BCI conditions as their favorite suggests a successful demonstration of our system. The comments ensuing from the open-ended interview questions suggest that one reason for this was that the system helped them achieve a level of musical improvisation that they would not have been able to otherwise, in other words, helped them be more creative. Another reason given was that they felt that their favorite BCI condition felt more responsive which suggests that the timing of the additions are crucial to preference. Adjectives participants used to describe their favorite condition included ‘‘*responsive*”, “*fun*”, “*enjoyed*”

"*dynamic*", "*fluid*", and "*creative*" which suggests that also, simply, the system was enjoyable to use and fun.

One of the important points brought up by musical experts in our pre-design interviews was that experienced musicians would not appreciate interference with their creative process. As noted, BrAAHMS was thus designed for beginner to intermediate pianists. Interestingly, one of our findings was that out of our group of beginner to intermediate players, the relatively more experienced users did show more variance in their preference for adaptations whereas less experienced ones showed a clear preference for the BCI conditions over the constant and non-adaptive conditions. In future musical brain-controlled systems, it would be prudent to consider and incorporate musical expertise and experience into the design process and the adaptation strategy.

While each of the two BCI conditions were preferred over either of the constant and non-adaptive conditions, the choice of which of the two BCI conditions was better varied across subjects. When designing our adaptations, we purposefully had not assumed whether to add musical additions when participants' cognitive workload was low and remove the additions when cognitive workload was high (BCIlow), or vice-versa (BCIhigh). We thus implemented both possible BCI conditions to explore user preference and to account for individual differences. This is a topic for further investigation to understand which adaptations are best for which users or conditions, or perhaps under what conditions to switch adaptations from one to the other.

5.7 Experiment 3 - Conclusion

I have demonstrated that it is possible to measure brain signals using fNIRS that correlate with high and low levels of cognitive difficulty while playing the piano. I then used this semantically meaningful brain data to build a real-time brain-controlled system that adapts musically to the user's cognitive state. I have carried out pilot studies through an iterative design process to determine which musical additions

would add the most pleasing harmonic enhancements to users' playing without compromising control over rhythm or tempo. Finally, I showed in an experiment that 15 out of 20 users preferred a brain-controlled condition, and I have discussed the possible reasons for this through post-experiment interview analysis and participant expertise data.

I suggest that BrAAHMS increases the communication bandwidth between the human and the musical instrument, responding to the user's cognitive state and providing appropriate musical additions just when they are needed, without requiring the user's effort or attention to control them.

Chapter 6

Experiment 4: Trust Evaluation System

This work is conditionally accepted by the ACM journal TOIT: Yuksel et al. (2016) Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Transactions on Internet Technology (TOIT) Special Section on Affect and Interaction in Agent-based Systems and Social Media* In Submission [231].

6.1 Experiment 4 - Introduction

It is not unlikely that, before long, intelligent agents will become prolific in the fields of learning/education and creativity. Pedagogical virtual agents and robotic agents are already being investigated as educational tutors, e.g., [44, 137, 182, 178]. Virtual agents are also becoming increasingly used in situations requiring creative abilities, e.g., [7, 13, 126, 33]. It is therefore fitting that we investigate what factors will encourage users to interact with such agents and start treating them as peers, to prepare the way for when learning/education and creativity will be strongly influenced, if not dominated by, intelligent, embodied agents.

Intelligent agents are playing an increasingly prominent and ubiquitous role in our lives. Examples of such smart personal agents include Microsoft's Cortana [133], Apple's Siri [9], or Google's Google Now [71]. Agents can collaborate with

users or perform tasks on users' behalf autonomously, gradually becoming more effective as they learn the users' interests, habits and preferences [118].

Maes [118] states that two main problems need to be solved when building software agents: 1) *Competence*: decisions of when to help the user and what to help them with, and 2) *Trust*: how we can guarantee that the user will trust the agent in making those decisions.

Humans have been shown to have "algorithm aversion" where they show greater intolerance to an algorithmic error than a human error [37]. While current artificial intelligence remains fallible, it is crucial that as researchers in user-agent interactions, we understand and develop the second problem area defined by [118]: how to build and maintain trust in the human-computer relationship even when the system inevitably makes errors.

In order to engender trust in intelligent agents, researchers have been turning to paradigms developed in social psychology that investigate human-to-human relationships and applying them to human-computer interaction [175, 18]. A very well established paradigm is that of the physical attractiveness stereotype where people attribute other positive personality traits onto attractive people [39]. A correlation between physical attractiveness and trustworthiness has been shown across several fields. Perhaps most interestingly, it has been demonstrated in acts of transgression or breaches of trust. Adults view more attractive children's transgressions less negatively and view less attractive children's offenses as more of an enduring dispositional fault [38]. Mock juror studies show that physically attractive defendants are found to be guilty less often and are given shorter sentences [48, 124]. Studies of judicial decisions in real courtrooms show that attractiveness is highly predictive of both minimum and maximum sentences [201, 202] and that judges exhibited a strong attractiveness-bias in the bails and fines that they set [46].

As software agents will become increasingly embodied in the future, it seems unlikely that unattractive intelligent assistants will become the norm. However, in this thesis, I investigate whether I could leverage the power of the physical attractiveness stereotype to cope with deficiencies in the machine. I investigate the effects

of reliability and attractiveness of embodied agents on users' trust.

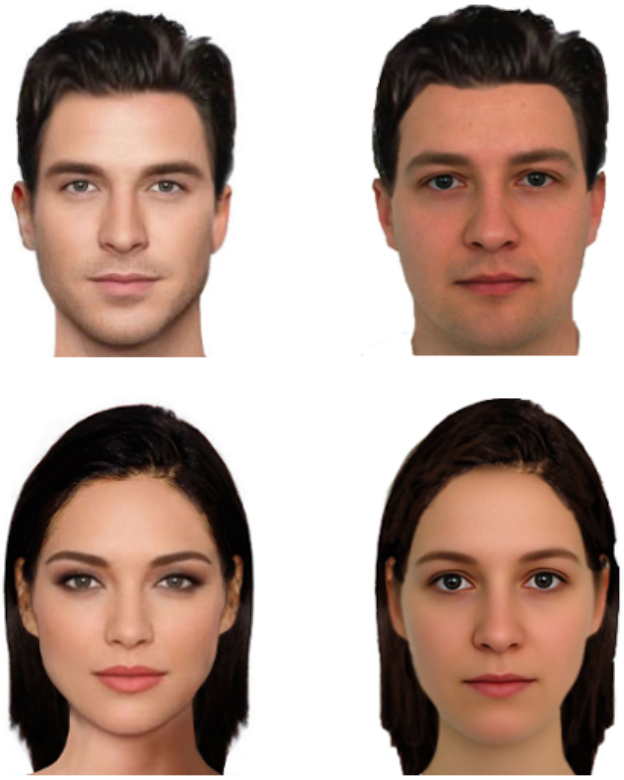


Figure 6.1: Highly attractive agents (left) courtesy of Dr. Stuart Gibson and Dr. Chris Solomon, University of Kent. Neutrally attractive agents (right) courtesy of Dr. Martin Gruendl, Beautycheck.

6.2 Hypothesis

In this paper I investigate two factors that affect trust: agent reliability and agent attractiveness. I test the following hypothesis:

Hypothesis: Reliability is more important than attractiveness in building trust in user-agent interactions.

6.3 Method

6.3.1 Materials

I chose to test highly attractive vs. neutrally attractive agents and avoid the unattractive condition as it would be highly unlikely that system designers would purposefully create unattractive agents. I prepared the material for agency appearance and voice accordingly.

6.3.1.1 Attractiveness of Physical Appearance

Research has shown that certain facial features are optimized for attractiveness, such as when the distance between the eyes and mouth is 36% of the face's length, or when the distance between the eyes is 46% of the face's width [156]. A large increase in the distance between the eyes and mouth is perceived as grotesque [189]. The eyes and mouth have consistently been the specific facial regions found to be most influential for facial attractiveness ratings [5].

I turned to two external sources that have worked in the area of facial features and attractiveness who kindly gave us their permission to use faces that they had already synthesized.

The two attractive agents' faces were provided courtesy of Dr. Stuart Gibson and Dr. Chris Solomon from the University of Kent. They created the faces as archetypal models of beauty from externally rated features of beauty by 100 people in the United Kingdom. Important features of beauty included distance between the eyes, nose length and width, thickness of lips and width of mouth. Figure 6.1 shows the highly attractive agents on the left.

The two neutrally attractive agents' faces were provided courtesy of Dr. Martin Gruendl at Beautycheck. The neutral female is averaged from 64 female faces. I was going to use a male face averaged together from male faces however the averaged male face was externally rated as more attractive (Dr. Martin Gruendl, Beautycheck) than the averaged female face, therefore I went with a more neutral looking male face. Figure 6.1 shows the neutrally attractive agents on the right.

I transplanted the same hair onto both the attractive and neutral faces as pilot studies showed that participants' ratings of agent attractiveness was being affected by differences in hair (Figure 6.1).

I then ran a preliminary study to verify the attractiveness of the agents. I asked 20 participants (11 female) to rate the physical attractiveness of the agents on a scale of 1-7 with 1 being 'Very Unattractive' and 7 being 'Very Attractive'. Participants rated their main gender of sexual preference (all participants were heterosexual so they rated the opposite gender's physical attractiveness). Figure 6.2 shows that participants rated the agents that were designed to be more physically attractive higher than those designed to be neutrally attractive.

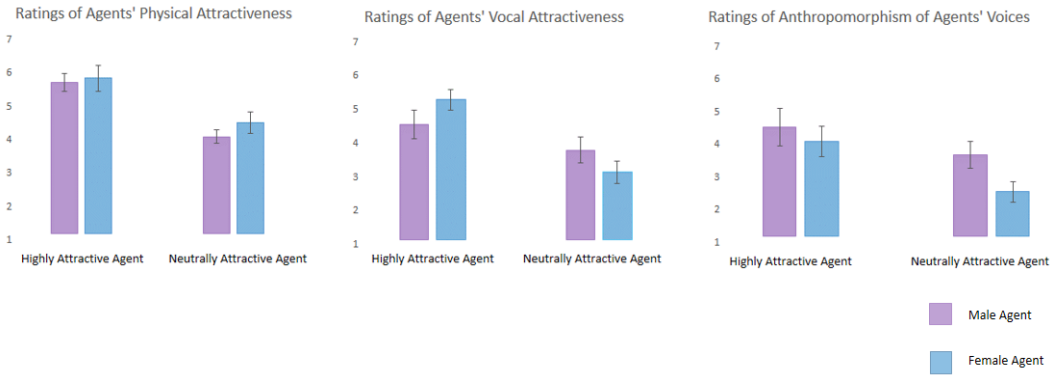


Figure 6.2: Mean and standard error of participants' ratings of the highly and neutrally attractive agents' *left*: physical attractiveness, *center:* vocal attractiveness, and *right:* vocal human-likeness.

6.3.1.2 Attractiveness of Voice

Vocal attractiveness has been shown to increase the attractiveness of face plus voice agency [236], it was therefore important that I maintained a differentiation of attractiveness in the vocalization of the agents. A stereotype for vocal attractiveness has been shown [236, 237] where attractive voices are rated more favorably in voice only and face plus voice conditions.

Studies from biological psychology have shown that men seem to prefer women with higher pitched voices which is associated with smaller body size [113], and women actually speak in a higher pitched voice to men they find attractive

[56]. Conversely, women seem to find men with lower pitched voices more attractive [29, 8] as this seems to indicate larger body size and other masculine traits [50].

I used these findings when building the voices of our agents using an internal Microsoft text to speech toolkit software. The female attractive voice was higher in pitch than the female neutral voice. The male attractive voice was lower in pitch than the male neutral voice. Furthermore, both the male and female attractive voices had prosodic contours with slight rises and falls in pitch, emphasis in places, and slight changes in the rhythm of speech. The neutral voices simply had a neutral tone throughout.

I asked 21 participants (11 female) to rate the vocal attractiveness of the voices on a scale of 1-7 with 1 being 'Very Unattractive' and 7 being 'Very Attractive'. Participants rated their main gender of sexual preference (all participants were heterosexual so they rated the opposite genders voices). Figure 6.2 shows that participants rated the voices that were designed to be more attractive higher than those designed to be neutrally attractive.

Due to the differences in prosodic contours creating more human-like speech patterns, I also asked participants to rate how human-like vs. machine-like the voices sounded, with 1 denoting 'Very Machine-Like' and 7 being 'Very Human-Like'. Figure 6.2 shows that participants thought that the neutral voices sounded more machine-like than the attractive voices. It is possible therefore, that this factor affects participants' preferences, which is considered when evaluating the results.

6.3.2 Experimental Design

Participants took part in a within-subject 2 (high vs. low reliability) x 2 (high vs. neutral attractiveness) agency design. Participants answered four sets of 15 questions. Each set had one of the four conditions:

- A highly reliable and highly attractive agent
  (Reliable_{HIGH} Attractive_{HIGH}).
- A highly reliable and neutrally attractive agent

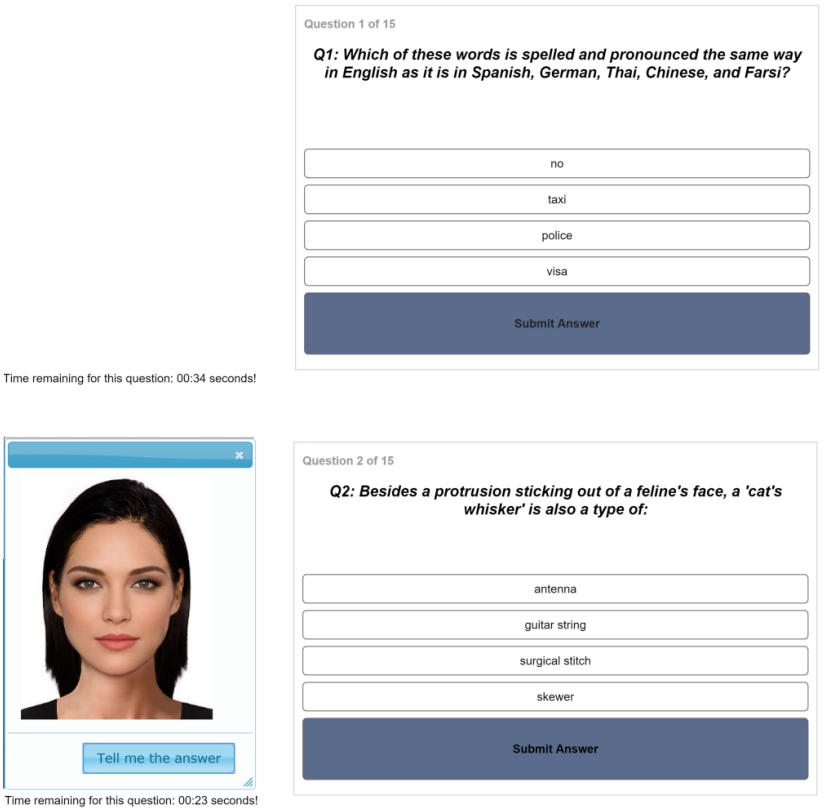


Figure 6.3: Top: Example of question without a suggestion. The time remaining can be seen on the bottom-left. Bottom: Example of a question with an agent offering a suggestion. Participants would hear the agent offering a suggestion as well. If they wished, the participant could hear the suggestion by pressing on the *'Tell me the answer'* button.

(Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$).

- A lesser reliable and highly attractive agent

  (Reliable$_{LOW}$ Attractive$_{HIGH}$).

- A lesser reliable and neutrally attractive agent

  (Reliable$_{LOW}$ Attractive$_{NEUTRAL}$)

The ordering of the conditions was alternated so that there was a fair distribution. I controlled for agents to be participants' main gender of sexual preference. All participants reported to be heterosexual and therefore only interacted with agents of the opposite gender.

In the high reliability conditions, 5 out of the six suggestions offered

by the agent were correct. In the low reliability conditions, 3 out of the six suggestions offered by the agent were correct. The percentages for high and low reliability (80% and 50% respectively) were taken from user trust literature [110]. The order of the correct and incorrect answers were randomized on each trial.

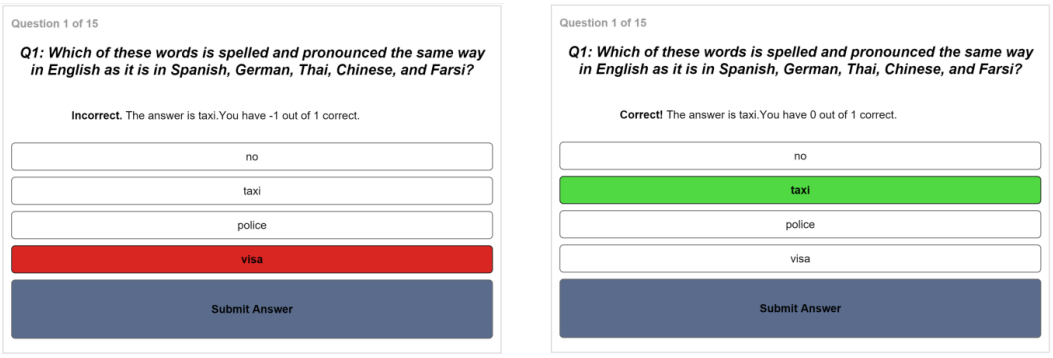


Figure 6.4: Examples of feedback for an incorrect (left) and correct (right) answer. Participants gained a point for every correct answer, and lost a point for every incorrect answer and every time they timed out in order to discourage guesswork.

The questions were multiple-choice and were general knowledge. Questions were on a broad range of topics and were designed to a difficulty level that required searching on most questions for most participants [80, 225]. Examples are included in Figure 6.3. Participants could search for the answers in a search engine on a separate monitor. They had 40 seconds to answer each question and they could see how much time they had left for each question with a countdown showing on the screen (Figure 6.3). For every question they answered correctly, they would gain a point, and they could move onto the next question. If they answered a question incorrectly they would lose a point (to discourage guesswork) and would have to keep attempting the question until they got the right answer or timed-out. If they timed-out they would also lose a point. They received a 2 minute break in between each set of questions. Participants were incentivized to reach high scores in the questions with extra gratuity. Participants always received immediate feedback when they answered a question (Figure 6.4).

Participants were told that every now and then an agent would appear on the screen to help them with the answer to a question. They were told that the agents

would offer a suggestion for the answer and that they could click on the '*Tell me the answer*' button to hear the suggestion (Figure 6.3). Before starting the experiment, participants were introduced to the two agents (highly and neutrally attractive) of their gender of preference from Figure 6.1.

For each set of 15 questions, 6 questions were picked pseudo-randomly on each trial to have suggestions. The suggestions were spread out so that the first five questions had 2 suggestions, the second five questions had 2 suggestions, and the last five questions had 2 suggestions.

6.3.2.1 Participants

Twenty participants took part in the study. One participant had to be excluded from the data set as they did not follow the instructions. Out of the remaining 19 (11 female), there was a wide range in ages: 17-51 years ($\mu = 30, \sigma = 11.24$). Ten participants were recruited internally from Microsoft Research, nine were recruited externally from the general public. Participants received a gift certificate for their participation. They were also offered an additional incentive of a $25 Amazon gift card if they could achieve 75% accuracy or higher in one of the sets of questions.

6.4 Mapping of Dependent Variables to Trust

Trust is a complex issue, and measuring it is no less simple. In this section, I present a mapping of how each dependent variable relates to the topic of trust in user-agent interactions.

6.4.1 Number of Questions With Button Presses

Users could choose to hear an agent's suggestion by pressing the 'Tell me the answer' button. I analyzed the number of questions with button presses to verify how often users opted to hear agents' suggestions. This is different from the number of button presses in total because in some trials users pressed the button more than once to re-listen to the suggestion for clarification or reminder purposes. There was no cost

to listening to agent suggestions.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

6.4.2 Self-reported Measures of Trust in Agent

This measure users' subjective ratings of trust in the agents.

6.4.3 Self-reported Measures of Perceived Accuracy of Agent

This measured users' perceived accuracy of agents. This reveals whether perceived agent reliability can be affected by other factors such as attractiveness, hence indicating where trust is being directed.

6.4.4 Number of Correct Suggestions Accepted

Due to the difference in the correct answers supplied by the agents with high versus low reliability, this measure should show this difference. A difference in the number of accepted correct answers between the high and low reliable agents will indicate a difference in trust between agents. Each correct answer gained the user a point.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

6.4.5 Number of Incorrect Suggestions Accepted

Similarly, this measure should show a difference between agents with high and low reliability due to the difference in incorrect answers supplied by those two groups of agents. A difference in the number of accepted correct answers between the high and low reliable agents will indicate a difference in trust between agents. This measure is different from number of correct suggestions accepted as users can select more than one incorrect answer for each question. Each incorrect answer cost the user a point.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

6.4.6 Length of Time Spent on Questions With Suggestions

This measure indicates trust by whether a user will accept the agent's suggestion or verify the answer with an Internet search. For the agents with high reliability, generally the longer the time taken, the less trust in the agent. However, due to agents with less reliability offering a greater number of incorrect suggestions, greater time spent may indicate that users were accepting incorrect suggestions, which is still indicative of trust. Therefore this measure is best compared between the agents with high and low reliability.

6.4.7 Number of Timeouts With Suggestions

Each question timed-out after 40 seconds, costing the user a point. Therefore, if a user timed-out on a question where an agent was offering a suggestion, this suggests a lack of trust. This is true across all conditions.

6.4.8 Number of Correct Answers

Agents with higher reliability offered a higher percentage of correct answers. Therefore this measure should show a difference between the agents with high and low reliability. A deviation from this would indicate a difference in trust between agents with high and low reliability.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

6.4.9 Number of Incorrect Answers

This measure should show an even greater difference between the agents with high and low reliability due to the fact that users can submit multiple incorrect answers per question. Therefore, questions with suggestions from an agent with low reliability can result in several incorrect answers.

This measure also acts as verification that users are not being exposed to skewed feedback from different conditions.

6.5 Results

6.5.1 Verification of Attractiveness Independent Variable

I first examined participants' ratings for the attractiveness of the agent's appearance and voice in order to verify the independent variables of attractiveness. Participants rated agents' physical and vocal attractiveness on two separate scales of 1-7 with 1 being 'Very Unattractive' and 7 denoting 'Very Attractive'.

6.5.1.1 Attractiveness of Physical Appearance

A paired t-test showed extremely significant different responses in participants' ratings of agents' physical attractiveness ($t(37) = 2.03, p = .00005$). Participants rated the agents designed to be highly attractive ($\mu = 4.8, \sigma = 0.8$) as significantly more physically attractive than the agents designed to be neutrally attractive ($\mu = 3.9, \sigma = 1.1$).

6.5.1.2 Attractiveness of Voice

A paired t-test also showed significant results for participants' ratings of agents' vocal attractiveness ($t(37) = 2.03, p = .02$). Participants rated the voices designed to be highly attractive ($\mu = 4.3, \sigma = 1.3$) as significantly more attractive than the voices designed to be neutrally attractive ($\mu = 3.9, \sigma = 1.2$).

These findings may be due, at least partially, to participants responding to the more human-like quality of the more attractive voices, as rated during preliminary studies of agents' voices (Figure 6.2).

6.5.2 Dependent Variables

6.5.2.1 Questionnaire Results on Trust and Accuracy

I carried out various 2 (high vs. low reliability) x 2 (high vs. neutral attractiveness) repeated measures (RM) ANOVAs with a Greenhouse-Geisser correction (in order to correct the degrees of freedom of the F-distribution to elicit a more accurate sig-

nificance value) across all dependent variables. For all significant results I calculated post-hoc analyses using Bonferroni corrections.

Results on *self-reported ratings of agent trust* showed a significant difference for the attractive agents ($F(1, 18) = 11.473, p = .003, \eta^2 = .389$) (Table 6.1). Post-hoc analysis revealed that users trusted the Reliable$_{HIGH}$Attractive$_{HIGH}$ agent significantly more than both of the neutrally attractive agents (including the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ ($p = .011$) and Reliable$_{LOW}$ Attractive$_{NEUTRAL}$ ($p = .005$) agents). Interestingly, there was no significant difference between the Reliable$_{LOW}$ Attractive$_{HIGH}$ agent and the highly reliable agents (Figure 6.5) (while Likert scales of 1-7 were given to participants with 1 denoting a low value and 7 denoting a high value, graphs show a scale of 0-6 to show better differences between conditions).

Results also showed *users' self-reporting of agent accuracy* was significantly different for more attractive agents ($F(1, 18) = 26.645, p < 0.001, \eta^2 = .510$) (Table 6.1). Post-hoc analysis showed that users thought both of the highly attractive agents were more accurate than the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent. Both the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent ($p = .005$) and even the Reliable$_{LOW}$ Attractive$_{HIGH}$ agent ($p = .006$) were significantly more accurate than the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent (Figure 6.5).

These results indicate that attractiveness is more important than reliability for user perceptions of trust and accuracy. I turn to data gathered from user-agent interaction to investigate these findings further.

6.5.2.2 User-Agent Interactions

There was a significant difference in the *number of correct suggestions accepted* for the reliability variable ($F(1, 18) = 25.352, p < 0.001, \eta^2 = .585$) and the interaction between reliability and attractiveness ($F(1, 18) = 4.569, p = .047$) (Table 6.1). As expected, post-hoc analysis showed that users accepted a significantly higher number of correct suggestions from the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent than for the lower reliability agents, regardless of attractiveness ($p < 0.001$). This was expected

Dependent Variable	Factor	F-value	p-value	η^2
Number of Questions With Button Presses	Reliability	3.485	.078	.162
Number of Questions With Button Presses	Attractiveness	3.547	.076	.165
Number of Questions With Button Presses	Interaction	4.455	.049	.198
Questionnaire on Trust	Reliability	0.548	.469	.030
Questionnaire on Trust	**Attractiveness**	9.642	**.006**	.389
Questionnaire on Trust	Interaction	0.986	.334	.052
Questionnaire on Accuracy	Reliability	1.536	.231	.079
Questionnaire on Accuracy	**Attractiveness**	18.731	**<.001**	.510
Questionnaire on Accuracy	Interaction	0.706	.412	.038
Number of Correct Suggestions Accepted	**Reliability**	25.352	**<.001**	.585
Number of Correct Suggestions Accepted	Attractiveness	3.888	.064	.178
Number of Correct Suggestions Accepted	Interaction	4.569	.047	.202
Number of Incorrect Suggestions Accepted	**Reliability**	65.681	**<.001**	.785
Number of Incorrect Suggestions Accepted	Attractiveness	0.029	.867	.002
Number of Incorrect Suggestions Accepted	Interaction	0.073	.790	.004
Time Spent on Questions With Suggestions	**Reliability**	13.027	**.002**	.420
Time Spent on Questions With Suggestions	**Attractiveness**	6.321	**.022**	.260
Time Spent on Questions With Suggestions	Interaction	0.073	.790	.004
Number of Timeouts With Suggestions	Reliability	6.892	.017	.277
Number of Timeouts With Suggestions	Attractiveness	1.000	.331	.053
Number of Timeouts With Suggestions	Interaction	.053	.821	.003
Number of Correct Answers	Reliability	3.512	.077	.163
Number of Correct Answers	Attractiveness	3.337	.084	.156
Number of Correct Answers	Interaction	.476	.499	.026
Number of Incorrect Answers	**Reliability**	5.773	**.027**	.243
Number of Incorrect Answers	Attractiveness	0.585	.454	.031
Number of Incorrect Answers	Interaction	.249	.624	.014

Table 6.1: Results: 2x2 RM-ANOVAS on agent reliability and attractiveness.
Note:All dependent variables that remained significant *after* the Bonferroni correction are highlighted in bold. All results have a df of (1,18).

as the highly reliable agents provided a higher number of correct suggestions in the first place than the agents with lesser reliability.

However, I did not expect for there to be *no significance* between the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent and the lower reliability agents (Figure B.10). This occurred even though the agents with lower reliability had many fewer correct suggestions to offer.

There was also a significant difference in the *number of incorrect suggestions accepted* for the reliability variable ($F(1, 18) = 65.681, p < 0.001, \eta^2 = .785$) (Table 6.1) as expected.

Results for the *time spent on questions when the agent offered suggestions* showed that there was a significant difference in both reliability ($F(1, 18) = 13.027, p = .002, \eta^2 = .420$) and attractiveness ($F(1, 18) = 6.321, p = .022, \eta^2 = .260$) (Table 6.1). Post-hoc analysis revealed that users spent significantly less time on questions with the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent than the Reliable$_{LOW}$ Attractive$_{NEUTRAL}$ ($p = .003$) agent (Figure 6.5). Interestingly, once again there were no significant difference between the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent and the low reliability agents. There were also no significant differences between the Reliable$_{LOW}$ Attractive$_{HIGH}$ agent and the highly reliability agents (Figure 6.5).

I used the median when measuring time as is standard practice when looking at reaction time distributions due to its robustness to skewness in the data. Figure 6.5 shows the mean of the medians.

There was a significant difference in the *number of incorrect answers* in reliability ($F(1, 18) = 5.773, p = .027, \eta^2 = .243$) (Table 6.1) as expected due to the difference in incorrect suggestions offered between agents with high and low reliability.

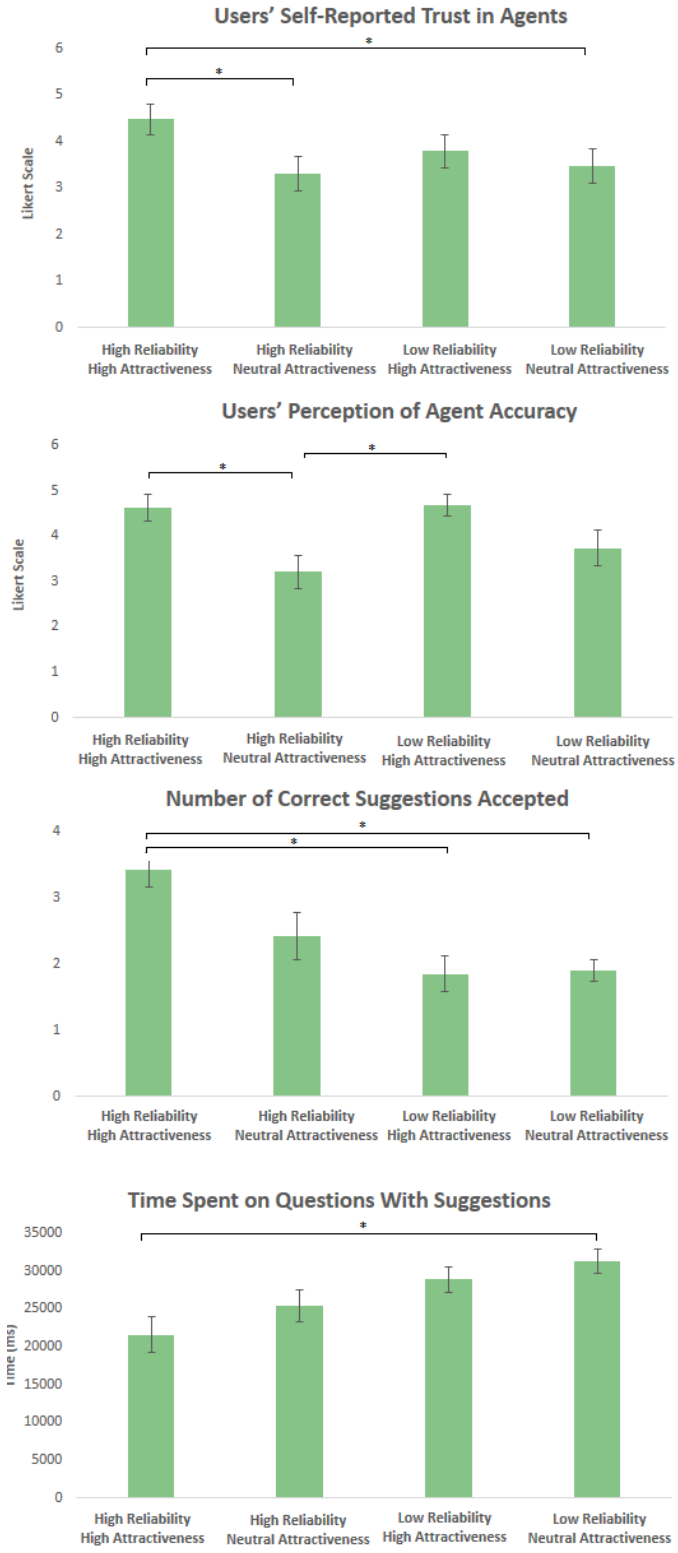


Figure 6.5: Mean and standard error of users' self-reported trust in agents, user perception of agent accuracy, number of correct suggestions accepted, and time spent on questions with suggestions.

However, there were no significant differences in the *number of correct answers* across conditions (Table 6.1).

There were also no significant differences in the *number of questions with button presses* (indicating the number of times users asked to hear the suggestion from the agent) (Table 6.1) which is not surprising considering there was no cost or penalty associated with hearing the suggestion. There were also no differences in the *number of timeouts in questions with suggestions* (Table 6.1).

I also examined all dependent variables for gender differences (bearing in mind all participants were heterosexual) with 2x2x2 RM-ANOVAs and found ***no differences in gender*** throughout the results.

6.6 Experiment 4 - Discussion

6.6.1 Reliability Manipulation Check

Users seem to trust agents with high attractiveness more and think that they are more accurate. I first verify that these results were not due to skewed exposure from different agents, by examining the dependent variables.

6.6.1.1 Number of Button Presses

There was no significant difference in the number of button presses to hear agents' suggestions between conditions (Table 6.1). Therefore users saw and heard the same number of suggestions across all conditions.

6.6.1.2 Number of Correct Suggestions Accepted

There was no significant difference in the number of correct suggestions accepted between the low reliability agents and the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent (Figure 6.5). Therefore differences in data between those three agents is not due to a skew in exposure to correct answers being accepted.

6.6.1.3 Number of Incorrect Suggestions Accepted

There was a significant difference in the number of incorrect suggestions accepted between the high and low reliability agents (Table 6.1) which was expected due to the higher number of incorrect suggestions provided by low reliability agents. Therefore any differences between highly and neutrally attractive agents were not due to skewed feedback in incorrect suggestions.

6.6.1.4 Number of Correct Answers

There was no significant difference in the number of correct answers across conditions, which once again verifies that the findings were not due to skewed exposure.

6.6.1.5 Number of Incorrect Answers

There was a significant difference in the number of incorrect answers between the high and low reliability agents (Table 6.1) which was to be expected due to higher number of incorrect suggestions in the low reliability agents. Therefore any differences between highly and neutrally attractive agents were not due to skewed feedback in incorrect answers.

It seems, therefore, that the results are due to users' own reactions to and perceptions of the agents. I now examine the data in terms of the hypotheses stated at the start.

6.6.2 Is Reliability More Important than Attractiveness in Building Trust in User-Agent Interactions?

I now address the hypothesis:
Reliability is more important than attractiveness in building trust in user-agent interactions.

Interestingly, results suggest a rejection of this hypothesis. Questionnaire data indicated users trusted the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent significantly

more than the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent. Users also found the Reliable $_{HIGH}$ Attractive$_{HIGH}$ agent to be significantly more accurate than the Reliable $_{HIGH}$Attractive$_{NEUTRAL}$ agent.This subjective evaluation of agents from users sheds light on how important agent attractiveness is in user-human interactions and relationships.

Objective user-agent interaction data showed us that the Reliable$_{HIGH}$ Attractive $_{NEUTRAL}$ agent was not treated differently from the agents with low reliability (Figure 6.5). While there were significant differences between the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent and the agents with low reliability in the number of correct suggestions accepted, there were no such significant differences between the Reliable$_{HIGH}$ Attractive $_{NEUTRAL}$ agent and the agents with low reliability. This was unusual because I had been expecting a significant difference in the number of correct suggestions accepted between the high and low reliability agents because the agents with lower reliability had fewer correct suggestions to offer.

This interaction data strongly suggests that users trusted the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent less than the Reliable$_{HIGH}$ Attractive$_{HIGH}$ agent by accepting fewer of their correct suggestions. It is also very interesting that users accepted the same number of correct suggestions from the agents with lower reliability as the Reliable$_{HIGH}$ Attractive$_{NEUTRAL}$ agent. Additionally, results showed that there was no significant difference in the length of time spent on questions between the Reliable$_{LOW}$ Attractive$_{HIGH}$ agent and the highly reliable agents. The time spent on considering answers reflects users' propensity to trust the agents' answers or to verify them by searching online. Together, these findings indicate a distinct lack of trust in user interactions with this highly reliable yet neutrally attractive agent. This result is really quite surprising and demonstrates how even a highly reliable agent's suggestions can be ignored by users if it is not accompanied with high levels of physical attractiveness.

Results indicate that it is not enough for an agent to simply be reliable for a user to trust it. It seems that **a reliable agent also needs to be attractive in order to gain users' trust enough for usage** at least within the boundaries

of our study. While attractiveness could be viewed as being just as important as reliability, the questionnaire data where users reported trusting attractive agents more and finding them more accurate suggests that it might even be more important than reliability.

These findings are representative of the complexity of human nature in interactions with software agents. They highlight the importance and value of acknowledging and exploring traits such as physical attractiveness. I recommend the leveraging of any trait that will help users trust software agents enough to engage with them, including physical attractiveness. This is not to deceive users, but instead, to encourage interaction. Otherwise, as demonstrated in these results, even highly reliable agents can be ignored by users.

6.6.3 Cognitive and Affective Trust

A possible explanation for greater user trust in the more attractive agents, even over the highly reliable but neutrally attractive agent, is the concept of affective trust over cognitive trust. Dowell et al. [2015] found that in the early phase of a relationship cycle in business (in the first 12 months), affective trust was especially important. In later, more mature phases of the relationship cycle (after the first 12 months), cognitive trust was a more important factor than before [45].

When there is less substantive evidence to build trust upon, it is thought that trust is based upon the hope or expectation that the task will be fulfilled, therefore cognitive trust would be less influential early on [45]. It is therefore possible that users were basing their trust on the agents in terms of how they felt about the agents (affective trust) which could have been influenced by the level of agent attractiveness. Lewis and Weigert [1985] suggest that the stronger the affective trust in relation to the cognitive trust, the less likely behavioral evidence to the contrary will weaken the relationship. This could have occurred with the agent with lesser reliability but higher attractiveness, and with the agent with high reliability but lesser attractiveness. If users had established an affective attitude towards them based on appearance, this could have created a blindness to evidence based agent

reliability.

Affective trust has been modeled into two sub-categories by Dowell et al. [2015]: 1. *Relational trust*, which is related to the 'leap of faith' aspect of affective trust where faith is placed in the other individual that they will act in a trustworthy way, and 2. *Intuitive trust* which is based on moods and feelings about the other individual. Dowell et al. [2015] found that relational trust was particularly important in the early phase of a relationship (which was considered as the first 12 months). The physical attractiveness stereotype has shown that people often attribute other positive traits onto physically attractive people, which could lend itself to this 'leap of faith' category.

It is important to note, however, that trust is a *mixture* of feeling and rational thinking [217]. In any user-agent interaction, it is not possible to exclude one or the other from the analysis of trust. However, affective trust seems to play a more important role than cognitive trust in primary group relations whereas cognitive trust seems to play a greater in secondary group relations [112]. Considering the deeply personal relationship most users have with their computers, it is foreseeable that an intelligent personal assistant or software agent will be included in the user's primary social group, thereby making the affective trust aspect of the relationship even more important. This would need to be investigated in a more long-term study of trust in user-agent interactions.

6.6.4 Physical versus Vocal Attractiveness in User-Agent Interactions

Is this study, I examined attractiveness as a combination of both physical and vocal attractiveness. It is possible that the two factors had an enhancing effect on the other. For example, vocal attractiveness has been shown to increase the attractiveness of face plus voice agency [236]. A separate study needs to be carried out to differentiate between the effectiveness of the two agencies.

After the completion of the main experiment, I carried out a pilot study to this effect to investigate the impact of vocal attractiveness *only* on engendering trust

in user-agent interactions. Four new participants (1 female) carried out the same experiment with only the agent's voice as stimuli. No visuals of the agents were presented.

Results showed no significant differences for any of the dependent variables across any conditions. The users also reported being so engrossed in the task that they could not tell a marked difference in the agents' voices. This was in contrast to previous findings where 10 participants who heard the agents' voices (also with no physical appearance) in an environment where they were not carrying out any tasks could clearly hear the differences in the voices (Figure 6.2).

These are preliminary results, however, which require further investigation to differentiate which agency factors elicit the attractiveness stereotype in humans.

6.6.5 Limitations

It may have been easier for users to accept the agents because they were merely 2D graphical representations of a real person and therefore avoided problems related to the uncanny valley. Systems with embodied virtual agents are currently being created that are able to move, gesture and lip-sync quite accurately when interacting with users (e.g., [215]). Examining the effects of reliability and attractiveness on the development of trust when agents are animated with natural gestures, facial expressions, and voice intonations would allow for a huge increase in agency parameters. This would provide a much more complex framework on which to design and investigate agent attractiveness and its effects on user-agent interactions.

The agents' interactions with users were merely in the form of question/answer types of interactions. While users carried out search scenarios and were motivated with money to take the tasks seriously, more real world, complex interaction between the user and the agent, when stakes are higher, would also be a ripe area for future exploration.

However, this work contributes to knowledge about agent design and its effects on humans, suggesting that attractiveness and reliability are both important considerations, and interact in ways that should be kept in mind in future user-agent

interactions.

6.7 Experiment 4 - Conclusion

I investigated the effects of reliability and attractiveness on the development of trust in user-agent interactions. I found that I had to reject our hypothesis that reliability was more important than attractiveness. Users reported trusting the highly attractive agents more and found them to be subjectively more accurate, even over the more reliable but less attractive agent. User-agent interaction data from the same number of correct suggestions accepted and length of time spent on questions showed that the highly reliable yet neutrally attractive agent was treated in the same way as the agents with low reliability. In addition, this agent was treated significantly differently from the agent who was highly reliable *and* highly attractive.

These results demonstrate that agent reliability is not simply more important than agent attractiveness in developing trust in user-agent interactions. In fact, attractiveness is just as important as reliability, and perhaps more so. These findings may be due to users having greater affective trust than cognitive trust in the highly attractive agents, especially as these were early-phase interactions when there is a less rational basis for trust.

I hope this thesis demonstrates that reliability by itself should not be the only point of focus in developing intelligent software agents. There has been a wealth of research conducted in social psychology for several decades highlighting the power and importance of the physical attractiveness stereotype. There has been further research demonstrating that humans respond to computers as if they were humans. Our study suggests that the attractiveness stereotype will follow humans into the future with our relationships and interactions with intelligent virtual agents.

Chapter 7

Conclusions

7.1 Summary of Work and Contributions

The next generation of implicit brain-computer interfaces provides the computer with more information about the human with no additional effort on the part of the user. This gives the computer the ability to respond intelligently to the human in real-time. However, such systems have been built for automated tasks. In this thesis, I built and evaluated two adaptive, implicit BCIs in two previously unexplored fields of application: ***learning*** and ***creativity***. I then built a trust evaluation system to explore what factors would encourage users to trust and use such intelligent systems.

In support of this thesis, I have made several contributions to the next generation of brain-computer interfaces and HCI research, as outlined below:

- In Chapter 3, I demonstrated that high and low cognitive workload could be measured using fNIRS when users play the piano during a **feasibility study**.
- In Chapter 4:

  - I addressed the usage of real-time, adaptive, implicit brain-computer interfaces in the **field of learning.**

  - I built an **adaptive, interactive musical learning system, BACh,** that guided learners' progress by measuring when they could cognitively handle

more information and provided them with the next stage of learning at the right moment.

- I demonstrated that users played the musical pieces with **increased speed and accuracy using BACh** than they normally would in a control condition.

- I demonstrated through questionnaire and interview data that participants felt they learnt the pieces better with BACh and they liked timings of the level changes.

- I implemented individual confidence threshold automation for additional classification accuracy of the brain signal rather than fixed percentage thresholds.

- I implemented difficulty level specific parameters to real-time data classification.

- In Chapter 5:

  - I addressed the usage of real-time, adaptive, implicit brain-computer interfaces in the **field of creativity.**

  - I built an **adaptive, intelligent, novel musical instrument system, BrAAHMS**, that used cognitive workload in real-time to add and remove musical harmonies during musical improvisation on the piano through a series of iterative design techniques.

  - I demonstrated through experimental evidence from interviews and subjective rankings that **users felt more creative** using BrAAHMS than a control condition because the system was responsive to them.

  - I implemented individual confidence threshold automation for additional classification accuracy of the brain signal rather than fixed percentage thresholds.

- In Chapter 6:

  - I built a **trust evaluation system** that explored factors that would affect user trust in intelligent systems.

- I investigated the factors of **virtual agent reliability and attractiveness** in a QA search task.

- I demonstrated through behavioural and questionnaire data that reliability alone is not enough for users to trust a virtual agent, and that **attractiveness is just as, if not more, important.**

7.2 Future Work

7.2.1 Future Work on Learning

BACh takes a first step in using cognitive workload, measured by brain sensing, as a determining factor in a user interface for adaptive learning. While meta-analyses have shown that fixed scheduling when increasing task difficulty is not helpful [219, 220], now that there is sufficient data for the length of time participants spent on each level, I can also investigate and compare other conditions such as using random intervals to increase difficulty level based on means or ranges of the times spent on the different levels in this study.

I can also look at increasing task difficulty based on judgment from an expert such as a piano teacher or self-judgment from the player of when to move onto the next level. The latter is very similar to the normal condition in this paper as learners were encouraged to learn how they normally would. They therefore segmented the piece between the left and right hands and increased the complexity as their competency grew.

The addition of emotion detection to BACh would also serve as a very powerful next step. If a learning system could detect both cognitive workload *and* affective state, it could be very potent learning tool indeed. There has been much work carried out on the detection of emotion using physiological sensing and facial expression recognition in the field of affective computing for several decades [22]. Such measurements could be used in conjunction with fNIRS or other brain sensing devices that are relatively resilient to motion artifacts. Emotion and learning are very closely tied together [166], with frustration often preceding giving up. The

role of emotion in learning has been described as a counter-clockwise movement through the four quadrants of the valence-arousal model of emotion by Kort et al. [97] (Figure 7.1). For example, the learner could start off in quadrant I where they are excited to learn a piece of music (or an instrument). They then might move into quadrant II as they hit a point of confusion or puzzlement, yet at this point they are still motivated to overcome the problem as they are in a state of high arousal, being in the top half of Figure 7.1 [97]. At some point it would not be uncommon for the musician to slip into the lower half of Figure 7.1 into quadrant III where there is negative valence in response to a self-perceived sense of failure. This section of the quadrant is the point at which it is most likely for the learner to give up, as frustration precedes discarding the learning attempt [166]. As the learner makes progress from this point by consolidating what works and what does not work, they move into quadrant IV [97].

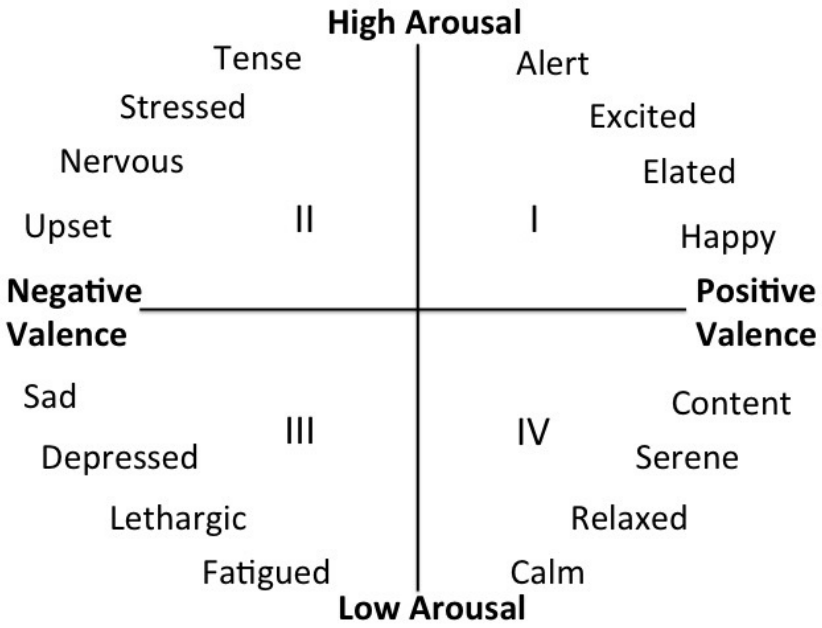


Figure 7.1: The valence-arousal model of emotion.

This movement through the valence-arousal model of emotion has yet to be explored in real-time with a musical interface and could provide a powerful tool both

by itself or in conjunction with cognitive workload. Just as cognitive workload can be measured passively in real-time, there are emerging technologies facial expression recognition and physiological sensing that make the measurement of emotion an non-invasive and passive occurrence.

Lastly, I address the topic of generalizability to other fields of learning. There has been a wealth of learning literature on other subjects such as mathematics or electrical engineering. The underlying premise of BACh is to increase the level of difficulty in a task when cognitive workload falls below a certain threshold measured by brain sensing. This can be applied, or at least investigated, in any field where information can be presented in increasing steps of difficulty.

7.2.2 Future Work on Creativity

Similar to the field of learning, the field of creativity would benefit enormously with the measurement of cognitive workload *and* user affective state in conjunction. For example, in the field of music, emotion is considered to be the essence, if not the purpose, of music [132]. Emotion could be used as a determining factor in directing musical improvisation or composition. However, much more generally, affect is considered vital to the development of creativity.

The NSF funded Workshop on creativity and creativity support tools discussed the importance of developing methods to measure the complex cognitive and affective processes associated with discovery and innovation that takes place over weeks and months in creative tasks [191, 192].

While BrAAHMS is a creative task that focuses on music, it carries implications for a broader reach into other fields of HCI where the user's goals are more non-specific/creative. In previous BCI work, there is generally a specific goal, where it is assumed that the user knows exactly what they want to do, such as find the most efficient UAV route [3] or direct a robot to its correct location [199]. In many tasks however there are occasions where we do not know exactly what our goal will look like until it is reached, such as in exploratory search or data analysis. In his seminal paper, Tukey [209] spoke of the importance of the non-specificity of goals in

scientific fields: "*Science - and engineering... does NOT begin with a tidy question. Nor does it end with a tidy answer.*" The work in this paper could be viewed as a demonstration of how semantically meaningful brain data can be used in a creative task, where the ultimate goal is non-specific, and the user does not know exactly what the outcome will look like but has a general idea of what they want [205].

Exploratory data analysis in the field of visual analytics is a prime candidate for such work. Analysts are faced with large and complex datasets that they must process, visualise, analyze, and subsequently extract the most critical information out of. In fact, to overcome some of these obstacles, Riche proposes and discusses the use of brain measurements including fNIRS [179]. It is a fertile field in which to explore the use of measuring cognitive state in non-specific goals. Thus, future work in this direction could be viewed in the broader context of helping users achieve non-specific goals with brain-based automation.

7.2.3 Future Work on Trust

The next step in agency in a wide range of areas such as education or entertainment will be the animation of intelligent agents with natural gestures and voice intonations. Systems with embodied and animated virtual agents are currently being created that are able to move, gesture and lip-sync quite accurately when interacting with users (e.g., [215]). In the future, as these systems become more mainstream, factors such as agency attractiveness will become even more important. Examining the effects of reliability and attractiveness on the development of trust when agents are animated with natural gestures, facial expressions, and voice intonations would allow for a huge increase in agency parameters. This would provide a much more complex framework on which to design and investigate agent attractiveness and its effects on user-agent interactions. However, such systems will have to overcome the uncanny valley if they are going to engender user trust, especially if attractiveness is a factor. The disturbing nature of the uncanny valley is likely to overpower any amount of attractiveness in the agent, however, this is an area that is yet to be explored.

In my system, interactions with users were merely in the form of question/answer types of interactions. While search scenarios were used and motivated participants with money to take the tasks seriously, more real world, complex interaction between the user and the agent, when stakes are higher, would also be a ripe area for future exploration.

However, this work contributes to knowledge about agent design and its effects on humans, suggesting that attractiveness and reliability are both important considerations, and interact in ways that should be kept in mind in future user-agent interactions.

7.3 Closing Remarks

This thesis demonstrates that it is possible to apply real-time, adaptive, implicit brain-computer interfaces to fields that are not just automation of tasks, but instead, more intelligent, personalized responses such as the fields of learning and creativity. Systems such as BACh and BrAAHMS clearly aid users learn with increased speed and accuracy or create something they could not have done by themselves, respectively.

As discussed in section 7.2.2, such systems could become even more potent by measuring user affective state in conjunction with cognitive workload. That will be the direction I will be taking in my own next chapter, post-thesis. In this way, we will continue to give computers more information about their users, so that they may continue to respond increasingly intelligently in return.

Bibliography

[1] D. A. Abwender and K. Hough. Interactive effects of characteristics of defendant and mock juror on U.S. participants' judgment and sentencing recommendations. *The Journal of social psychology*, 141(5):603–615, 2001.

[2] E. D. Adrian and B. Mathews. The Berger Rhythm: Potential Changes from the Occipital Lobes in Man. *BRAIN*, 57:355–385, 1934.

[3] D. Afergan, E. M. Peck, E. T. Solovey, A. Jenkins, S. W. Hincks, E. T. Brown, R. Chang, and R. J. Jacob. Dynamic difficulty using brain metrics of workload. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 3797–3806, 2014.

[4] D. Afergan, T. Shibata, S. W. Hincks, E. M. Peck, B. F. Yuksel, R. Chang, and R. J. Jacob. Brain-Based Target Expansion. *Proc. UIST 2014*, 2014.

[5] T. R. Alley and K. A. Hildebrandt. *Social and applied aspects of perceiving faces*, chapter Determinants and consequences of facial aesthetics, pages 101–140. Lawrence Erlbaum Associates, Hillsdale, NJ., 1988.

[6] B. Z. Allison and J. A. Pineda. Effects of soa and flash pattern manipulations on erps, performance, and preference: implications for a bci system. *International journal of psychophysiology*, 59(2):127–140, 2006.

[7] J. Anstey and D. Pape. Scripting the interactor: An approach to vr drama. In *Proceedings of the 4th conference on Creativity & cognition*, pages 150–156. ACM, 2002.

[8] C. Apicella, D. Feinberg, and F. Marlowe. Voice pitch predicts reproductive success in male hunter-gatherers. *Biology Letters*, 3(6):682–684, 2007.

[9] 2005. `http://www.apple.com/ios/siri/`.

[10] B. Arslan, A. Brouse, J. Castet, R. Léhembre, C. Simon, J.-J. Filatriau, and Q. Noirhomme. A real time music synthesis environment driven with biological signals. In *Proc. IEEE ICASSP 2006*, 2006.

[11] D. W. Ash and D. H. Holding. Backward versus forward chaining in the acquisition of a keyboard skill. *Human Factors*, 32:139–146, 1990.

[12] H. Ayaz, P. A. Shewokis, S. Bunce, K. Izzetoglu, B. Willems, and B. Onaral. Optical brain monitoring for operator training and mental workload assessment. *Neuroimage*, 59(1):36–47, 2012.

[13] R. Aylett and S. Louchart. I contain multitudes: creativity and emergent narrative. In *Proceedings of the 9th ACM Conference on Creativity & Cognition*, pages 337–340. ACM, 2013.

[14] M. J. Baker and G. A. Churchill Jr. The impact of physically attractive models on advertising evaluations. *Journal of Marketing research*, pages 538–555, 1977.

[15] M. H. Beauchamp, A. Dagher, J. A. D. Aston, and J. Doyon. Dynamic functional changes associated with cognitive skill learning of an adapted version of the Tower of London task. *NeuroImage*, 20:1649–1660, 2003.

[16] E. Berscheid and E. Hatfield. Interpersonal attraction. 1969.

[17] E. R. Bialo and J. Sivin-Kachala. The effectiveness of technology in schools: A summary of recent research. 1996.

[18] T. W. Bickmore and R. Picard. Establishing and Maintaining Long-Term Human-Computer Relationships. 12(2):293–327, 2005.

[19] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Botchoubey, A. Kubler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398:297–298, 1999.

[20] R. Brünken, J. L. Plass, and D. Leutner. Direct Measurement of Cognitive Load in Multimedia Learning. *Educational Psychologist*, 38(1):53–61, 2003.

[21] P. W. Burgess, A. Quayle, and C. D. Frith. Brain regions involved in prospective memory as determined by positron emission tomography. *Neuropsychologia*, 39(6):545–555, 2001.

[22] R. A. Calvo, S. D'Mello, J. Gratch, and A. K. (Eds). *The Oxford Handbook of Affective Computing.* Oxford University Press, 2014., 2014.

[23] S. Chaiken. Communicator physical attractiveness and persuasion. *Journal of Personality and social Psychology*, 37(8):1387, 1979.

[24] B. Chance, E. Anday, S. Nioka, S. Zhou, L. Hong, K. Worden, C. Li, T. Murray, Y. Ovetsky, D. Pidikiti, et al. A novel method for fast imaging of brain function, non-invasively, with light. *Optics express*, 2(10):411–423, 1998.

[25] C. Chang and C. Lin. Speeded near infrared spectroscopy (NIRS) response detection. *PLOS ONE*, 5:11, 2010.

[26] C. Chang and C. Lin. LIBSVM: a library for support vector machines. *Transactions on Intelligent Systems and Technology*, 10:1–39, 2011.

[27] Y. C. D. Chew and E. Caspary. Museegk: a brain computer musical interface. In *Extended Abstracts CHI 2011*, pages 1417–1422, 2011.

[28] M. T. Chi, R. Glaser, and E. Rees. Expertise in problem solving. Technical report, DTIC Document, 1981.

[29] S. A. Collins. Men's voices and women's choices. *Animal behaviour*, 60(6):773–780, 2000.

[30] E. Cutrell and D. Tan. Bci for passive input in hci. ACM Press.

[31] L. A. Dabbish and R. S. Baker. Administrative assistants as interruption mediators. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*, pages 1020–1021. ACM, 2003.

[32] D. Damos and C. D. Wickens. The identification and transfer of time-sharing skills. *Acta Psychologica*, 46:15–39, 1980.

[33] N. Davis, C.-P. Hsiao, K. Yashraj Singh, L. Li, and B. Magerko. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 196–207. ACM, 2016.

[34] T. de Jong. Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38:105–134, 2010.

[35] M. Deppe, W. Schwindt, H. Kugel, H. Plassmann, and P. Kenning. Nonlinear responses within the medial prefrontal cortex reveal when specific implicit information influences econo.pdf. *Journal of Neuroimaging*, pages 171–182, 2005.

[36] M. D'Esposito, B. R. Postle, and B. Rypma. Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. *Experimental brain research*, 133(1):3–11, 2000.

[37] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. 143(6):1–13, 2014.

[38] K. Dion. Physical Attractiveness and Evaluation of Children ' S Transgressions *. *Journal of Personality and Social Psychology*, 24(2):207–213, 1972.

[39] K. Dion, E. Berscheid, and E. Walster. What is beautiful is good. *Journal of personality and social psychology*, 24(3):285–290, 1972.

[40] K. K. Dion. Young children's stereotyping of facial attractiveness. *Developmental Psychology*, 9(2):183, 1973.

[41] K. K. Dion and E. Berscheid. Physical attractiveness and peer perception among children. *Sociometry*, pages 1–12, 1974.

[42] K. K. Dion and S. Stein. Physical attractiveness and interpersonal influence. *Journal of Experimental Social Psychology*, 14(1):97–108, 1978.

[43] E. Donchin, K. M. Spencer, and R. Wijesinghe. The mental prosthesis: assessing the speed of a p300-based brain-computer interface. *Rehabilitation Engineering, IEEE Transactions on*, 8(2):174–179, 2000.

[44] J. T. Doswell. It's virtually pedagogical: pedagogical agents in mixed reality learning environments. In *ACM SIGGRAPH 2005 Educators program*, page 25. ACM, 2005.

[45] D. Dowell, M. Morrison, and T. Heffernan. The changing importance of affective trust and cognitive trust across the relationship lifecycle: A study of business-to-business relationships. *Industrial Marketing Management*, 44:119–130, 2015.

[46] A. C. Downs and P. M. Lyons. Natural observations of the links between attractiveness and initial legal judgments. *Personality and Social Psychology Bulletin*, 17(5):541–547, 1991.

[47] A. H. Eagly, R. D. Ashmore, M. G. Makhijani, and L. C. Longo. What is beautiful is good, but: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological bulletin*, 110(1):109, 1991.

[48] M. G. Efran. The Effect of Guilt , of Physical Appearance on the Judgment Severity Interpersonal Attraction , and of Recommended Punishment in a Simulated. *Journal of Research in Personality*, 8:45–54, 1974.

[49] G. P. Elovitz and J. Salvia. Attractiveness as a biasing factor in the judgments of school psychologists. *Journal of School Psychology*, 20(4):339–345, 1983.

[50] S. Evans, N. Neave, and D. Wakelin. Relationships between vocal characteristics and body size and shape in human males: an evolutionary explanation for a deep male voice. *Biological psychology*, 72(2):160–163, 2006.

[51] S. H. Fairclough. Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2):133–145, Jan. 2009.

[52] L. A. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical Neurophysiology*, 70(6):510–523, 1988.

[53] A. Feingold. Good-looking people are not what we think. *Psychological bulletin*, 111(2):304, 1992.

[54] B. J. Fogg and C. Nass. Silicon sycophants: the effects of computers that flatter. *International Journal of Human-Computer Studies*, 46(5):551–561, 1997.

[55] P. T. Fox, M. E. Raichle, M. A. Mintun, and C. Dence. Glucose During Focal Physiologic Neural Activity Nonoxidaive Consumption. *Science*, 241:462–464, 1988.

[56] P. Fraccaro, B. Jones, J. Vukovic, F. Smith, C. Watkins, D. Feinberg, A. Little, and L. Debruine. Experimental evidence that women speak in a higher voice pitch to men they find attractive. *Journal of Evolutionary Psychology*, 9(1):57–67, 2011.

[57] T. K. Fredericks, S. D. Choi, J. Hart, S. E. Butt, and A. Mital. An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads. *International Journal of Industrial Ergonomics*, 35:1097–1107, 2005.

[58] H. H. Friedman and I. C. Friedman. Whom do students trust? *Journal of Communication*, 26(2):48–49, 1976.

[59] H. H. Friedman, M. J. Santeramo, and A. Traina. Correlates of trustworthiness for celebrities. *Journal of the Academy of Marketing Science*, 6(4):291–299, 1978.

[60] O. Friman, I. Volosyak, and A. Gräser. Multiple channel detection of steady-state visual evoked potentials for brain-computer interfaces. *Biomedical Engineering, IEEE Transactions on*, 54(4):742–750, 2007.

[61] F. Galán, M. Nuttin, E. Lew, P. W. Ferrez, G. Vanacker, J. Philips, and J. d. R. Millán. A brain-actuated wheelchair: asynchronous and non-invasive brain–computer interfaces for continuous control of robots. *Clinical Neurophysiology*, 119(9):2159–2169, 2008.

[62] M. Galizio and C. Hendrick. Effect of musical accompaniment on attitude: The guitar as a prop for persuasion. *Journal of Applied Social Psychology*, 2(4):350–359, 1972.

[63] H. Garavan, D. Kelley, A. Rosen, S. M. Rao, and E. A. Stein. Practice-related functional activation changes in a working memory task. *Microscopy research and technique*, 51(1):54–63, 2000.

[64] A. Gevins and M. E. Smith. Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral cortex*, 10(9):829–39, Sept. 2000.

[65] A. Gevins, M. E. Smith, H. Leong, L. McEvoy, S. Whitfield, R. Du, and G. Rush. Monitoring working memory load during computer-based tasks with eeg pattern recognition methods. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 40(1):79–91, 1998.

[66] A. Gevins, M. E. Smith, L. McEvoy, and D. Yu. High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, 7:374–385, 1997.

[67] A. Girouard, E. Solovey, and R. Jacob. Designing a passive brain computer interface using real time classification of functional near-infrared spectroscopy. *International Journal of Autonomous and Adaptive Communications Systems*, 6(1):26–44, 2013.

[68] A. Girouard, E. T. Solovey, L. M. Hirshfield, K. Chauncey, A. Sassaroli, S. Fantini, and R. J. Jacob. Distinguishing difficulty levels with non-invasive brain activity measurements. In *Human-Computer Interaction–INTERACT 2009*, pages 440–452. Springer, 2009.

[69] A. Glass, D. L. McGuinness, and M. Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236, 2008.

[70] J. A. Gliner, G. A. Morgan, and N. L. Leech. *Research methods in applied settings: An integrated approach to design and analysis*. Psychology Press, 2000.

[71] 2005. `https://www.google.com/landing/now/`.

[72] Grierson, M. Composing with brainwaves: minimal trial P300b recognition as an indication of subjective preference for the control of a musical instrument. *ICMC*, 2008.

[73] Grierson, M., Kiefer, C., and Yee-King, M. Progress Report on the EAVI BCI Toolkit for Music: Musical Applications of Algorithms for use with consumer brain computer interfaces. *ICMC*.

[74] W. Griffitt and R. Veitch. Preacquaintance attitude similarity and attraction revisited: Ten days in a fall-out shelter. *Sociometry*, pages 163–173, 1974.

[75] A. Gundel and G. F. Wilson. Topographical changes in the ongoing eeg related to the difficulty of mental tasks. *Brain topography*, 5(1):17–25, 1992.

[76] J. Harnum. *Basic Music Theory: How to Read, Write, and Understand Written Music*. Sol-UT Press, 2009.

[77] J. Harrison, K. Izzetoglu, H. Ayaz, B. Willems, S. Hah, U. Ahlstrom, H. Woo, P. Shewokis, S. C. Bunce, and B. Onaral. Cognitive workload and learning assessment during the implementation of a next-generation air traffic control technology using functional near-infrared spectroscopy. *IEEE Transactions on Human-Machine Systems*, 44(4):429–440, 2014.

[78] L. M. Hirshfield, R. Gulotta, S. Hirshfield, S. Hincks, M. Russell, R. Ward, T. Williams, and R. Jacob. This is your brain on interfaces: enhancing usability testing with functional near-infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 373–382. ACM, 2011.

[79] M. Hosoda, E. F. Stone-Romero, and G. Coats. The effects of physical attractiveness on job-related outcomes: A meta-analysis of experimental studies. *Personnel Psychology*, 56(2):431, 2003.

[80] Hyperion. *Who wants to be a millionaire*. Hyperion, NYC, NY, USA, 2000.

[81] S. T. Iqbal, X. S. Zheng, and B. P. Bailey. Task-evoked pupillary response to mental workload in human-computer interaction. *Extended abstracts of the 2004 conference on Human factors and computing systems CHI 04*, page 1477, 2004.

[82] K. Izzetoglu, H. Ayaz, J. Menda, M. Izzetoglu, A. Merzagora, P. A. Shewokis, K. Pourrezaei, and B. Onaral. Applications of functional near infrared imaging: case study on uav ground controller. In *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*, pages 608–617. Springer, 2011.

[83] L. A. Jackson, J. E. Hunter, and C. N. Hodge. Physical attractiveness and intellectual competence: A meta-analytic review. *Social Psychology Quterly*, pages 108–122, 1995.

[84] C.-L. C. K. James A. Kulik and R. L. Bangert-Drowns. Effectiveness of computer-based education in elementary schools. *Computers in Human Behavior*, 1:59–74, 1985.

[85] M. M. Johnson and J. P. Rosenfeld. Oddball-evoked p300-based method of deception detection in the laboratory ii: Utilization of non-selective activation of relevant knowledge. *International Journal of Psychophysiology*, 12(3):289–306, 1992.

[86] C. Johnson-George and W. C. Swap. Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of Personality and Social Psychology*, 43(6):1306, 1982.

[87] J. Jonides, E. E. Smith, R. A. Koeppe, E. Awh, S. Minoshima, and M. A. Mintun. Spatial working-memory in humans as revealed by pet. 1993.

[88] S. Kalyuga. The expertise reversal effect. *Educational psychologist*, 38(1):23–31, 2003.

[89] S. Kalyuga. Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19(4):509–539, 2007.

[90] S. Kalyuga. Cognitive Load Theory: How Many Types of Load Does It Really Need? *Educational Psychology Review*, 23:1–19, 2011.

[91] D. B. Kandel. Similarity in real-life adolescent friendship pairs. *Journal of personality and social psychology*, 36(3):306, 1978.

[92] C. Kelly and H. Garavan. Human functional neuroimaging of brain changes associated with practice. *Cerebral Cortex*, 15(August):1089–1102, 2005.

[93] D. A. Kenny and W. Nasby. Splitting the reciprocity correlation. *Journal of Personality and Social Psychology*, 38(2):249, 1980.

[94] W. Klimesch. Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research reviews*, 29(2):169–195, 1999.

[95] R. B. Knapp and H. S. Lusted. A bioelectric controller for computer music applications. *Computer music journal*, pages 42–47, 1990.

[96] E. Koechlin, G. Basso, P. Pietrini, S. Panzer, and J. Grafman. The role of the anterior prefrontal cortex in human cognition. *Nature*, 399:148–51, May 1999.

[97] B. Kort, R. Reilly, and R. W. Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *ICALT*, page 0043. IEEE, 2001.

[98] A. Kostov and M. Polak. Parallel man-machine training in development of eeg-based cursor control. *IEEE Transactions on Rehabilitation Engineering*, 8(2):203–205, 2000.

[99] J. K. Kroger, F. W. Sabb, C. L. Fales, S. Y. Bookheimer, M. S. Cohen, and K. J. Holyoak. Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cerebral cortex*, 12(5):477–485, 2002.

[100] J. A. Kulik. *Technology Assessment in Education and Training*, chapter Meta-analytic Studies of findings on computer-based instruction., pages 9–34. Eva Baker and Harold F. O'Neil Jr, 1994.

[101] T. M. M. Lahtinen, J. P. Koskelo, T. Laitinen, and T. K. Leino. Heart rate and performance during combat missions in a flight simulator. *Aviation Space and Environmental Medicine*, 78(4):387–391, 2007.

[102] J. H. Langlois, L. A. Roggman, R. J. Casey, J. M. Ritter, L. A. Rieser-Danner, and V. Y. Jenkins. Infant preferences for attractive faces: Rudiments of a stereotype? *Developmental psychology*, 23(3):363, 1987.

[103] J. H. Langlois and C. W. Stephan. Beauty and the beast: The role of physical attractiveness in the development of peer relations and social behavior. *Developmental social psychology: Theory and research*, pages 152–168, 1981.

[104] S. Le Groux, J. Manzolli, and P. F. Verschure. Disembodied and Collaborative Musical Interaction in the Multimodal Brain Orchestra. *Proc. NIME*, pages 309–314, 2010.

[105] E. J. Lee. Flattery may get computers somewhere, sometimes: The moderating role of output modality, computer gender, and user gender. *International Journal of Human Computer Studies*, 66(11):789–800, 2008.

[106] E. J. Lee. I like you, but I won't listen to you: Effects of rationality on affective and behavioral responses to computers that flatter. *International Journal of Human Computer Studies*, 67(8):628–638, 2009.

[107] J. Lee, J. Kim, and J. Y. Moon. What makes internet users visit cyber stores again? key design factors for customer loyalty. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 305–312, 2000.

[108] J. C. Lee and D. S. Tan. Using a low-cost electroencephalograph for task classification in HCI research. *Proc. UIST 2006*, 2006.

[109] R. Leeb, C. Keinrath, D. Friedman, C. Guger, R. Scherer, N. C., M. Garau, A. Antley, A. Steed, M. Slater, and G. Pfurtscheller. Walking by thinking: The brainwaves are crucial, not the muscles! *Presence*, 15(5):500–514, 2006.

[110] S. LeeTiernan, E. Cutrell, M. Czerwinski, and H. Hoffman. Effective notification systems depend on user trust. In *Proceedings of Human-Computer Interaction–Interact*, pages 54–69, 2001.

[111] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal. The persona effect: affective impact of animated pedagogical agents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 359–366. ACM, 1997.

[112] J. D. Lewis and A. Weigert. Trust as a social reality. *Social forces*, 63(4):967–985, 1985.

[113] X. Liu and Y. Xu. What makes a female voice attractive? *Proceedings of the XVIIth International Congress of Phonetic Sciences*, (August):2174–1277, 2011.

[114] A. Lucier. Statement on: music for solo performer. In D. Rosenboom, editor, *Biofeedback and the Arts, Results of Early Experiments.*, pages 60–61. Vancouver: Aesthetic Research Center of Canada Publications, 1976.

[115] K. E. Lukanov, H. A. Maior, and M. L. Wilson. Using fnirs in usability testing: understanding the effect of web form layout on mental workload. 2016.

[116] n. . . p. . . t. . C. v. . O. y. . . Lusted, H. S. and Knapp, R. B., journal = Scientific American.

[117] D. M. Mackie and L. T. Worth. Feeling good, but not thinking straight: The impact of positive mood on persuasion. *Emotion and social judgments*, 23:210–219, 1991.

[118] P. Maes. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, 1994.

[119] H. A. Maior, M. Pike, S. Sharples, and M. L. Wilson. Examining the reliability of using fnirs in realistic hci settings for spatial and verbal tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3039–3042. ACM, 2015.

[120] M. Maltz and J. Meyer. Cue utilization in a visually demanding task. In *Proceedings of the Human Factors and Ergonomics Society... Annual Meeting*, volume 1, page 283. Human Factors and Ergonomics Society, 2000.

[121] D. S. Manoach, G. Schlaug, B. Siewert, D. G. Darby, B. M. Bly, A. Benfield, R. R. Edelman, and S. Warach. Prefrontal cortex fMRI signal changes are correlated with working memory load. *Neuroreport*, 8(2):545–549, 1997.

[122] R. E. Mayer. Fifty years of creativity research. In R. J. Sternberg, editor, *Handbook of Creativity*. 1999.

[123] R. E. Mayer. Cognitive theory of multimedia learning. 2005.

[124] R. Mazzella and A. Feingold. The effects of physical attractiveness, race, socioeconomic status, and gender of defendants and victims on judgments of mock jurors: A meta-analysis1. *Journal of Applied Social Psychology*, 24(15):1315–1338, 1994.

[125] S. M. McClure, J. Li, D. Tomlin, K. S. Cypert, L. M. Montague, and P. R. Montague. Neural correlates of behavioral preference for culturally familiar drinks. *Neuron*, 44(2):379–87, Oct. 2004.

[126] J. McCormick, A. Nash, S. Hutchison, K. Vincs, S. Nahavandi, and D. Creighton. Recognition: combining human interaction and a digital performing agent. In *Proceedings of the 2014 Virtual Reality International Conference*, page 19. ACM, 2014.

[127] D. J. McFarland, W. A. Sarnacki, G. Townsend, T. Vaughan, and J. R. Wolpaw. The p300-based brain–computer interface (bci): effects of stimulus rate. *Clinical Neurophysiology*, 122(4):731–737, 2011.

[128] W. J. McGuire. The nature of attitudes and attitude change. *The handbook of social psychology*, 3(2):136–314, 1969.

[129] G. R. McMillan, G. Calhoun, M. Middendorf, J. Schnurer, D. Ingle, and V. Nasman. Direct brain interface utilizing self-regulation of steady-state visual evoked response (ssver). In *Proc. RESNA95 Annu. Conf*, volume 15, pages 693–695, 1995.

[130] S. Mealla, A. Väljamäe, M. Bosi, and S. Jordà. Listening to your brain: Implicit interaction in collaborative music performances. *Proc. NIME '11*, 2011.

[131] J. Mellinger, G. Schalk, C. Braun, H. Preissl, W. Rosenstiel, N. Birbaumer, and A. Kubler. An meg-based brain-computer interface (bci). *Neuroimage*, 36(3):581–593, 2007.

[132] L. B. Meyer. *Emotion and meaning in music*. University of chicago Press, 2008.

[133] 2005. `http://www.microsoft.com/en-us/mobile/experiences/cortana/`.

[134] J. R. Millan, F. Renkens, J. Mouriño, and W. Gerstner. Noninvasive brain-actuated control of a mobile robot by human eeg. *Biomedical Engineering, IEEE Transactions on*, 51(6):1026–1033, 2004.

[135] A. G. Miller. Role of physical attractiveness in impression formation. *Psychonomic Science*, 19(4):241–243, 1970.

[136] J. Mills and E. Aronson. Opinion change as a function of the communicator's attractiveness and desire to influence. *Journal of Personality and Social Psychology*, 1(2):173, 1965.

[137] M. Milne, D. Powers, and R. Leibbrandt. Development of a software-based social tutor for children with autism spectrum disorders. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7*, pages 265–268. ACM, 2009.

[138] E. Miranda and A. Brouse. Toward direct brain-computer musical interfaces. *Proc. NIME*, pages 216–219, 2005.

[139] E. R. Miranda. Brain-computer music interface for composition and performance. *Int J Disabil Hum Dev*, 5(2):119–126, 2006.

[140] E. R. Miranda. Plymouth brain-computer music interfacing project: from eeg audio mixers to composition informed by cognitive neuroscience. *Intl. Journal of Arts and Technology*, 3(2/3):154–176, 2010.

[141] E. R. Miranda, K. Sharman, K. Kilborn, and A. Duncan. On Harnessing the Electroencephalogram for the Musical Braincap. *Computer Music Journal*, 27(2):80–102, 2003.

[142] S. Moghimi, A. Kushki, S. Power, A. M. Guerguerian, and T. Chau. Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy. *Journal of neural engineering*, 9(2):026022, 2012.

[143] Y. Moon. *Intimate self-disclosure exhanges: Using computers to build reciprocal relationships with consumers.* Division of Research, Harvard Business School, 1998.

[144] R. Moreno. Cognitive load theory: more food for thought. *Instructional Science*, 38(November 2009):135–141, 2010.

[145] J. Morkes, H. K. Kernal, and C. Nass. Humor in task-oriented computer-mediated communication and human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 215–216. ACM, 1998.

[146] C. Nass and Y. Moon. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1):81–103, Jan. 2000.

[147] C. Neuper, R. Scherer, M. Reiner, and G. Pfurtscheller. Imagery of motor actions: Differential effects of kinesthetic and visual–motor mode of imagery in single-trial eeg. *Cognitive Brain Research*, 25(3):668–677, 2005.

[148] S. A. Nida and J. E. Williams. Sex-stereotyped traits, physical attractiveness, and interpersonal attraction. *Psychological reports*, 41(3f):1311–1322, 1977.

[149] R. Nkambou, R. Mizoguchi, and J. Bourdeau. Advances in intelligent tutoring systems. 2010.

[150] L. A. Nordholm. Beautiful patients are good patients: evidence for the physical attractiveness stereotype in first impressions of patients. *Social Science & Medicine. Part A: Medical Psychology & Medical Sociology*, 14(1):81–83, 1980.

[151] H. S. Nwana. Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, 4(4):251–277, 1990.

[152] Z. Obrenovic. A flexible system for creating music while interacting with the computer. *Proc. MM 2005*, pages 996–1004, 2005.

[153] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. V. Gerven. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1):63–71, 2003.

[154] F. Paas and J. J. G. van Merrienboer. Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive load approach. *Journal of Educational Psychology*, 86:122–133, 1994a.

[155] F. G. Paas. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of educational psychology*, 84(4):429, 1992.

[156] P. M. Pallett, S. Link, and K. Lee. New "golden" ratios for facial beauty. *Vision Research*, 50(2):149–154, 2010.

[157] G. L. Patzer. Source credibility as a function of communicator physical attractiveness. *Journal of business research*, 11(2):229–241, 1983.

[158] M. P. Paulus and L. R. Frank. Ventromedial prefrontal cortex activation is critical for preference judgments. *Neuroreport*, 14(10):1311–5, July 2003.

[159] E. M. Peck, D. Afergan, and R. J. K. Jacob. Investigation of fNIRS brain sensing as input to information filtering systems. *Proc. Augmented Human 2013*, pages 142–149, 2013.

[160] E. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang. Using fNIRS Brain Sensing to Evaluate Information Visualization Interfaces. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 473–482, 2013.

[161] Peck, E. M., Afergan, D., Yuksel, B. F., Lalooses, F., Jacob, R.J.K. Using fNIRS to measure mental workload in the real world. In Fairclough, S.H. and Gilleade, K., editor, *Advances in physiological computing.* 2014.

[162] S. E. Petersen, H. Van Mier, J. A. Fiez, and M. E. Raichle. The effects of practice on the functional anatomy of task performance. *Proceedings of the National Academy of Sciences*, 95(3):853–860, 1998.

[163] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. L. Da Silva. Mu rhythm (de) synchronization and eeg single-trial classification of different motor imagery tasks. *NeuroImage*, 31(1):153–159, 2006.

[164] G. Pfurtscheller and C. Neuper. Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134, 2001.

[165] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlogl, B. Obermaier, and M. Pregenzer. Current trends in graz brain-computer interface (bci) research. *IEEE Transactions on Rehabilitation Engineering*, 8(2):216–219, 1997.

[166] R. Picard. *Affective computing*, 252, 1997.

[167] T. W. Picton. The p300 wave of the human event-related potential. *Journal of clinical neurophysiology*, 9(4):456–479, 1992.

[168] M. F. Pike, H. a. Maior, M. Porcheron, S. C. Sharples, and M. L. Wilson. New York, New York, USA.

[169] W. Piston and M. DeVoto. *Harmony.* 1987.

[170] J. Platt et al. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

[171] R. A. Poldrack. Imaging brain plasticity: conceptual and methodological issues a theoretical review. *Neuroimage*, 12(1):1–13, 2000.

[172] L. J. Prinzel, F. G. Freeman, M. W. Scerbo, P. J. Mikulka, and A. T. Pope. Effects of a Psychophysiological System for Adaptive Automation on Performance, Workload, and the Event-Related Potential P300 Component. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(4):601–613, 2003.

[173] 2015. `http://ethw.org/Oral-History:Ralph_Benjamin#Information_Theory.3B_digital_plot_transmission_and_mouse_patent`.

[174] N. Ramnani and A. M. Owen. Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience*, 5(3):184–194, 2004.

[175] B. Reeves and C. Nass. *The Media Equation: How people treat computers, television, and new media like real people and places.* CLSI and Cambridge university press, Cambridge, 1996.

[176] J. K. Rempel, J. G. Holmes, and M. P. Zanna. Trust in close relationships. *Journal of personality and social psychology*, 49(1):95, 1985.

[177] G. Repovs and A. Baddeley. The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience*, 139(1):5–21, 2006.

[178] D. Richards, M. J. Jacobson, M. Taylor, A. Newstead, C. Taylor, J. Porte, I. Kelaiah, and N. Hanna. Learning to be scientists via a virtual field trip. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pages 1463–1464. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[179] N. Riche et al. Beyond system logging: human logging for evaluating information visualization. In *Proceedings of the BELIV 2010 Workshop*, 2010.

[180] D. Rosenboom. The Performing Brain. *Computer Music Journal*, 14:48–66, 1990.

[181] M. D. Rugg, P. C. Fletcher, K. Allan, C. D. Frith, R. Frackowiak, and R. Dolan. Neural correlates of memory retrieval during recognition memory and cued recall. *Neuroimage*, 8(3):262–273, 1998.

[182] M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse. Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1613–1622. ACM, 2010.

[183] R. J. Salden, F. Paas, N. J. Broers, and J. J. van Merrinboer. Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training. *Instructional Science*, 32(1-2):153–172, 2004.

[184] C. A. Samuels and R. Ewy. Aesthetic perception of faces during infancy*. *British Journal of Developmental Psychology*, 3(3):221–228, 1985.

[185] F. Sauvet, J. C. Jouanin, C. Langrume, P. Van Beers, Y. Papelier, and C. Dussault. Heart rate variability in novice pilots during and after a multi-leg cross-country flight. *Aviation, space, and environmental medicine*, 80(10):862–869, 2009.

[186] W. Schnotz and C. Kürschner. A eeconsideration of cognitive load theory. *Educational Psychology Review*, 19:469–508, 2007.

[187] H. Schultheis and A. Jameson. Assessing cognitive load in adaptive hypermedia systems: Physiological and behavioral methods. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3137:225–234, 2004.

[188] N. Schwarz, H. Bless, and G. Bohner. Mood and persuasion: Affective states influence the processing of persuasive communications. *Advances in experimental social psychology*, 24:161–199, 1991.

[189] J. H. Searcy and J. C. Bartlett. Inversion and processing of component and spatial–relational information in faces. *Journal of Experimental Psychology: Human Perception and Performance*, 22(4):904, 1996.

[190] B. Shneiderman. Creativity Support Tools: A Grand Challenge for HCI Researchers. *Engineering the User Interface*, pages 1–9, 2009.

[191] B. Shneiderman, G. Fischer, M. Czerwinski, B. Myers, and M. Resnik. Creativity Support Tools : A Workshop Sponsored by the National Science Foundation, 2005.

[192] B. Shneiderman, G. Fischer, B. Myers, E. Edmonds, M. Eisenberg, and P. Jennings. Creativity Support Tools : Report from a U.S . National Science Foundation Sponsored Workshop. *Internationl Journal of Human-Computer Interaction*, 20(2):61–77, 2006.

[193] C. L. Sholbs. Improvement in type-writing machines, July 14 1868. US Patent 79,868.

[194] H. Sigall and N. Ostrove. Beautiful but dangerous: Effects of offender attractiveness and nature of the crime on juridic judgment. *Journal of Personality and Social Psychology*, 31(3):410, 1975.

[195] H. Sigall, R. Page, and A. C. Brown. The effects of physical attractiveness and evaluation on effort expenditure and work output. In *Proceedings of the Annual Convention of the American Psychological Association*, 1969.

[196] E. J. Sirevaag, A. F. Kramer, M. G. Coles, and E. Donchin. Resource reciprocity: An event-related brain potentials analysis. *Acta Psychologica*, 70(1):77–97, 1989.

[197] E. T. Solovey, A. Girouard, K. Chauncey, L. M. Hirshfield, A. Sassaroli, F. Zheng, S. Fantini, and R. J. K. Jacob. Using fNIRS Brain Sensing in Realistic HCI Settings : Experiments and Guidelines. 2009.

[198] E. T. Solovey, F. Lalooses, K. Chauncey, D. Weaver, M. Scheutz, A. Sassaroli, S. Fantini, P. Schermerhorn, and R. J. K. Jacob. Sensing Cognitive Multitasking for a Brain-Based Adaptive User Interface. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2011.

[199] E. T. Solovey, P. Schermerhorn, M. Scheutz, A. Sassaroli, S. Fantini, and R. J. K. Jacob. Brainput : Enhancing Interactive Systems with Streaming fNIRS Brain Input. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2012.

[200] N. K. Squires, K. C. Squires, and S. A. Hillyard. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and clinical neurophysiology*, 38(4):387–401, 1975.

[201] J. E. Stewart. Defendant's attractiveness as a factor in the outcome of criminal trials: An observational study1. *Journal of Applied Social Psychology*, 10(4):348–361, 1980.

[202] J. E. Stewart. Appearance and Punishment: The Attraction-Leniency Effect in the Courtroom. *The Journal of Social Psychology*, 125(3):373–378, 1985.

[203] J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.

[204] J. Sweller. *Instructional Design in Technical Areas*. ACER Press, 1999.

[205] J. Sweller and M. Levine. Effects of goal specificity on means–ends analysis and learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 1982.

[206] D. Szafir and B. Mutlu. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 11–20. ACM, 2012.

[207] D. Szafir and B. Mutlu. Artful: adaptive review technology for flipped learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1001–1010. ACM, 2013.

[208] E. Treacy Solovey, D. Afergan, E. M. Peck, S. W. Hincks, and R. J. Jacob. Designing implicit interfaces for physiological computing: Guidelines and lessons learned using fnirs. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(6):35, 2015.

[209] J. W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25, 1980.

[210] P. Ullsperger, G. Freude, and U. Erdmann. Auditory probe sensitivity to mental workload changes–an event-related potential study. *International Journal of Psychophysiology*, 40(3):201–209, 2001.

[211] Vamvakousis, Z., and Ramirez, R. P300 Harmonies: A Brain-Computer Musical Interface. *ICMC*, 2014.

[212] P. W. M. Van Gerven, F. Paas, J. J. G. Van Merriënboer, and H. G. Schmidt. Memory load and the cognitive pupillary response in aging. *Psychophysiology*, 41:167–174, 2004.

[213] A. Villringer and B. Chance. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neurosciences*, 20, 1997.

[214] L. Vygotsky. Interaction between learning and development. *Mind and Society*, pages 79–91, 1978.

[215] L. Wang, W. Han, F. K. Soong, and Q. Huo. Text driven 3d photo-realistic talking head. In *INTERSPEECH*, pages 3307–3308, 2011.

[216] A. Waytz, J. Heafner, and N. Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52:113–117, 2014.

[217] A. J. Weigert. *Sociology of everyday life*. Longman, 1981.

[218] N. Weiskopf, K. Mathiak, S. Bock, F. Scharnowski, R. Veit, W. Grodd, R. Goebel, and N. Birbaumer. Principles of a brain-computer interface (bci) based on real-time functional magnetic resonance imaging (fmri). *IEEE Transactions on Biomedical Engineering*, 51(6):966–970, 2004.

[219] C. D. Wickens, S. Hutchins, T. Carolan, and J. Cumming. Effectiveness of Part-Task Training and Increasing-Difficulty Training Strategies: A Meta-Analysis Approach. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2013.

[220] D. C. Wightman and G. Lintern. Part-task training for tracking and manual control. *Human Factors*, 27(3):267–283, 1985.

[221] G. F. Wilson and C. A. Russell. Performance Enhancement in an Uninhabited Air Vehicle Task Using Psychophysiologically Determined Adaptive Aiding. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(6):1005–1018, Dec. 2007.

[222] J. Wolpaw, Birbaumer, W. Heetderks, D. McFarland, P. Peckham, G. Schalk, E. Donchin, L. Quatrano, C. Robinson, and T. Vaughan. Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering*, 8(2):164–173, 2000.

[223] J. Wolpaw and D. McFarland. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. volume 101, pages 17849–17854, 2004.

[224] J. Wolpaw, D. McFarland, G. Neat, and C. Forneris. An eeg-based brain-computer interface for cursor control. *Electroencephalogram and Clinical Neurophysiology*, 78:252–258, 1991.

[225] S. Wood and K. Kovalchik. $Mental_{f}lossTrivia$: $Briskrefreshingfactswithouttheicecream.Puzzlewright, NYC, NY, USA$, 2012.

[226] D. Wu, C. Li, Y. Yin, C. Zhou, and D. Yao. Music composition from the brain signal: representing the mental state by music. *Computational intelligence and neuroscience*, 2010:267671, 2010.

[227] B. Xie and G. Salvendy. Prediction of mental workload in single and multiple task environments. *International journal of cognitive ergonomics*, 4:213–242, 2000.

[228] S. Yoo, T. Fairneny, N. Chen, S. Choo, L. Panych, H. Park, S. Lee, and F. Jolesz. Brain-computer interface using fmri: spatial navigation by thoughts. *Neuroreport*, 15(10):1591–1595, 2004.

[229] L. C. Young and G. S. Albaum. *Developing a Measure of Trust in Retail Relationships: A Direct Selling Application.* School of Marketing, University of Technology, Sydney, 2002.

[230] B. F. Yuksel, D. Afergan, E. M. Peck, G. Griffin, L. Harrison, N. Chen, R. Chang, and R. J. Jacob. BRAAHMS: A Novel Adaptive Musical Interface Based on Users' Cognitive State. *New Interfaces for Musical Expression (NIME)*, pages 136–139, 2015.

[231] B. F. Yuksel, P. Collisson, and M. Czerwinski. Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Transactions on Internet Technology (TOIT) Special Section on Affect and Interaction in Agent-based Systems and Social Media.*, (In Submission), 2016.

[232] B. F. Yuksel, M. Donnerer, J. Tompkin, and A. Steed. A novel brain-computer interface using a multi-touch surface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 855–858. ACM, 2010.

[233] B. F. Yuksel, K. B. Oleson, L. Harrison, E. M. Peck, D. Afergan, R. Chang, and R. J. Jacob. Learn Piano with BACh: An Adaptive Learning Interface that Adjusts Task Difficulty based on Brain State. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (2016)*, volume 2858388, pages 5372–5384, 2016.

[234] T. O. Zander, J. Brönstrup, R. Lorenz, and L. R. Krol. Towards bci-based implicit control in human–computer interaction. In *Advances in Physiological Computing*, pages 67–90. Springer, 2014.

[235] T. O. Zander and C. Kothe. Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *Journal of Neural Engineering*, 8(2), 2011.

[236] M. Zuckerman and R. E. Driver. What sounds beautiful is good: The vocal attractiveness stereotype. *Journal of Nonverbal Behavior*, 13(2):67–82, 1989.

[237] M. Zuckerman, H. Hodgins, and K. Miyake. The vocal attractiveness stereotype: Replication and elaboration. *Journal of Nonverbal Behavior*, 14(2):97–112, 1990.

Appendix A

Musical Pieces Used During the fNIRS Training Task.

The fifteen easy pieces presented to participants in random order for 30 seconds during the training task of Experiments 1, 2, and 3:

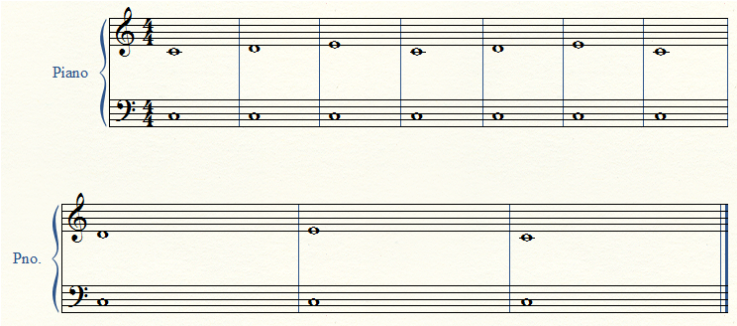


Figure A.1: Easy Piece 1

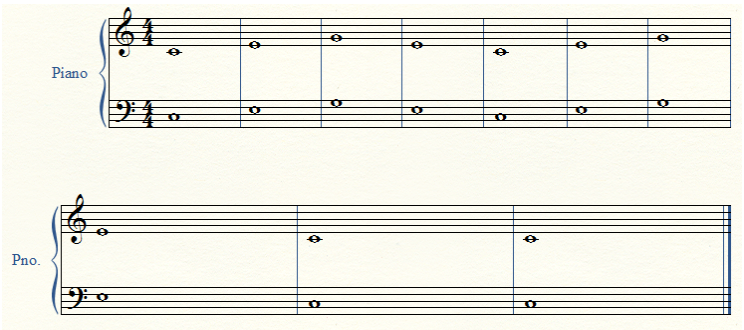


Figure A.2: Easy Piece 2

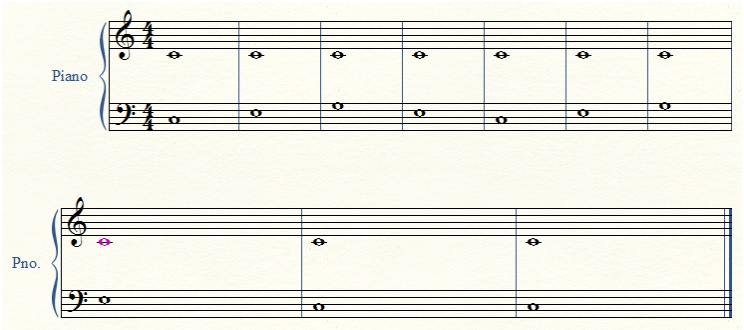


Figure A.3: Easy Piece 3

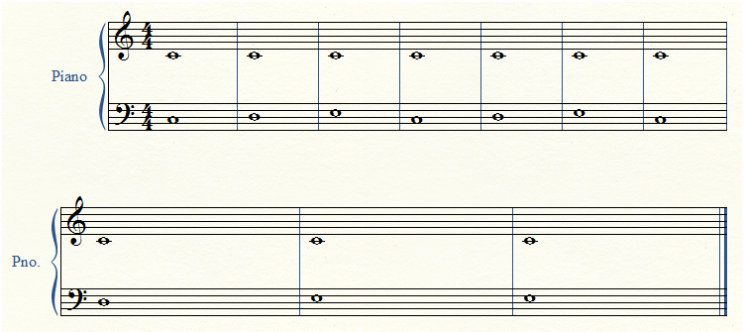


Figure A.4: Easy Piece 4

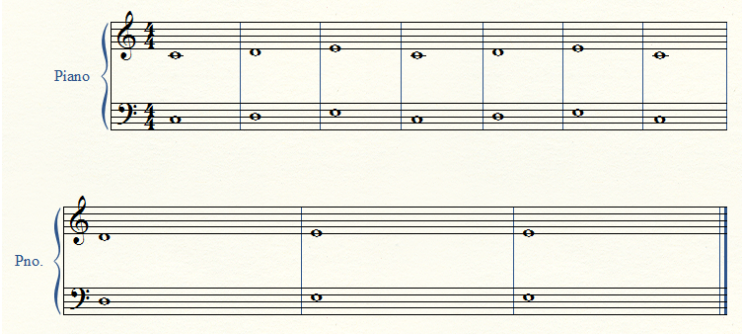


Figure A.5: Easy Piece 5

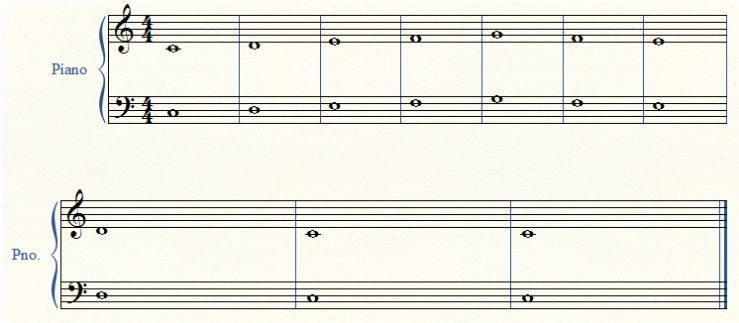


Figure A.6: Easy Piece 6

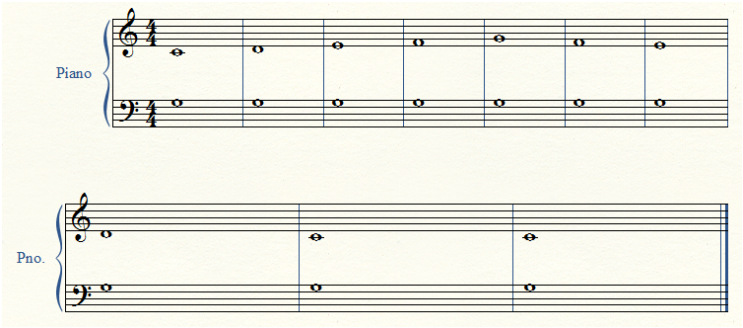


Figure A.7: Easy Piece 7

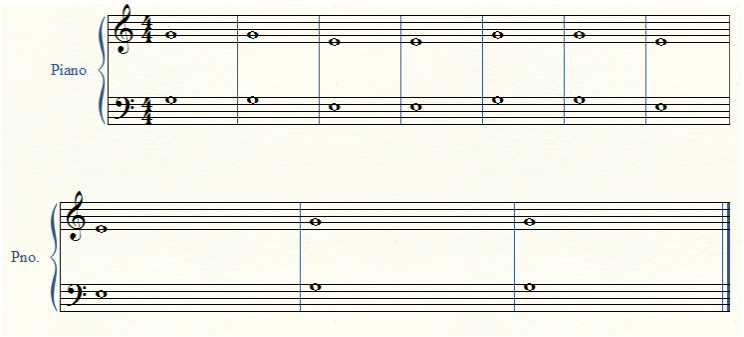


Figure A.8: Easy Piece 8

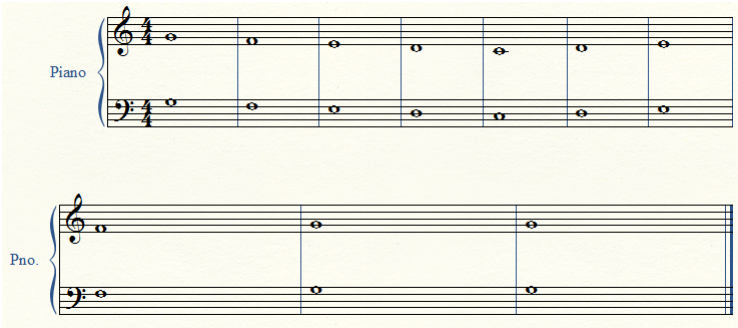


Figure A.9: Easy Piece 9

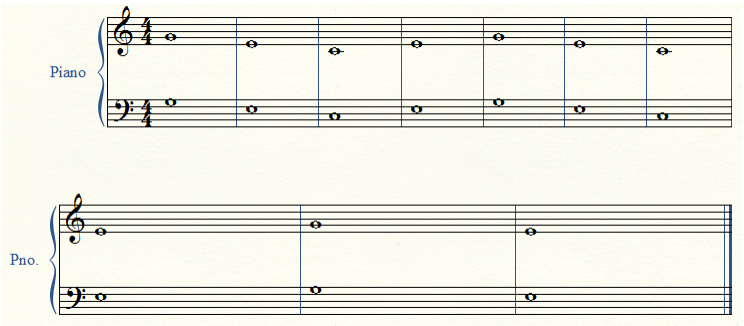


Figure A.10: Easy Piece 10

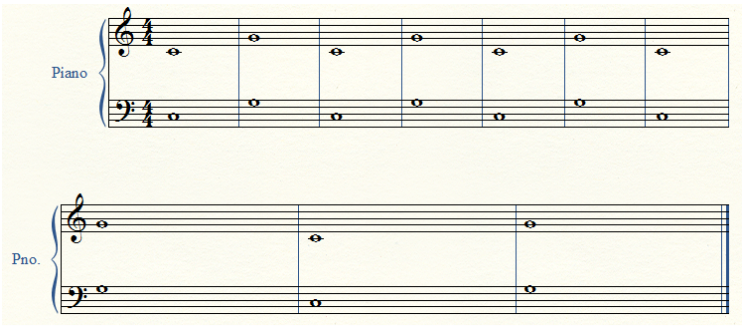


Figure A.11: Easy Piece 11

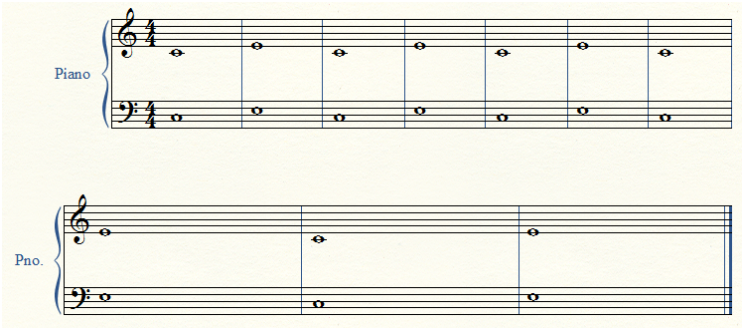


Figure A.12: Easy Piece 12

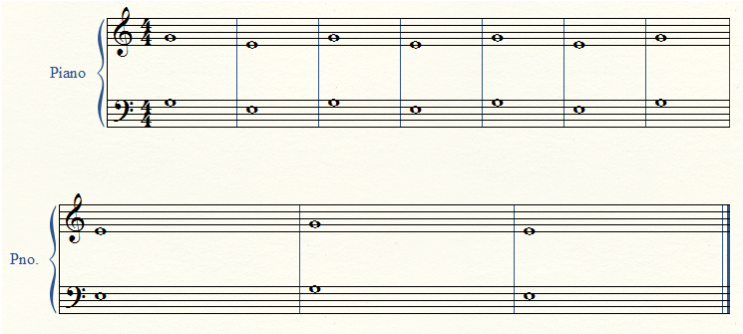


Figure A.13: Easy Piece 13

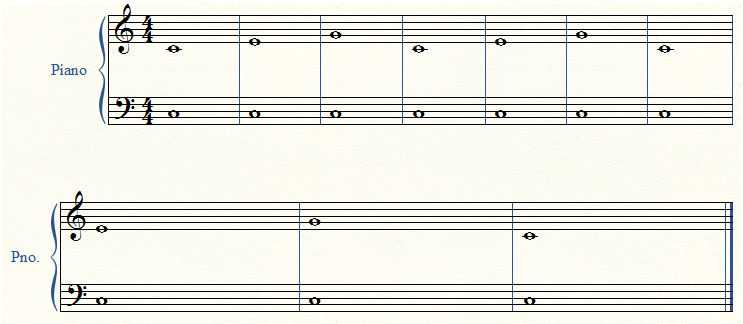


Figure A.14: Easy Piece 14

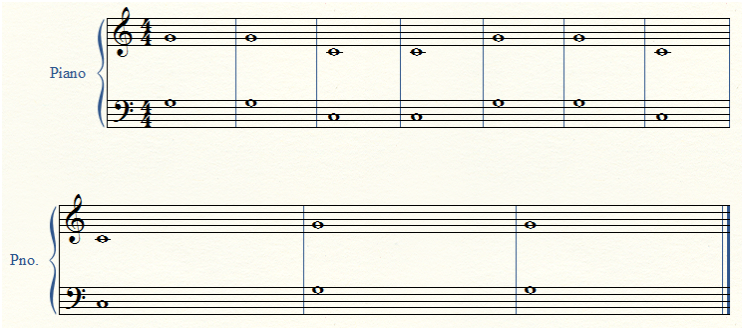


Figure A.15: Easy Piece 15

The fifteen hard pieces presented to participants in random order for 30 seconds during the training task of Experiments 1, 2, and 3:

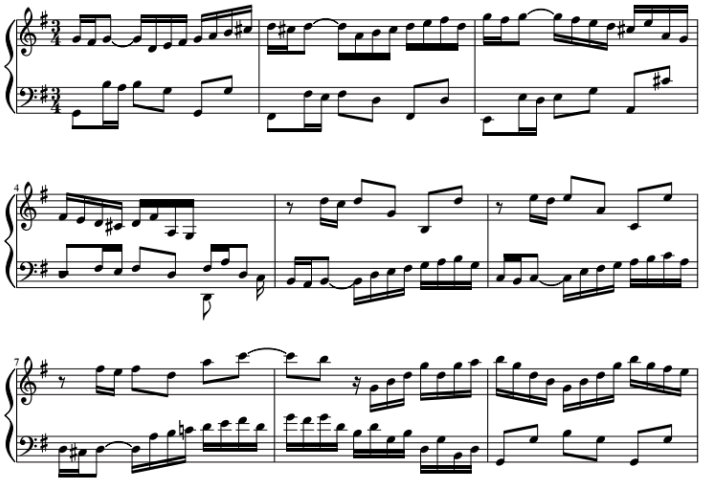


Figure A.16: Hard Piece 1

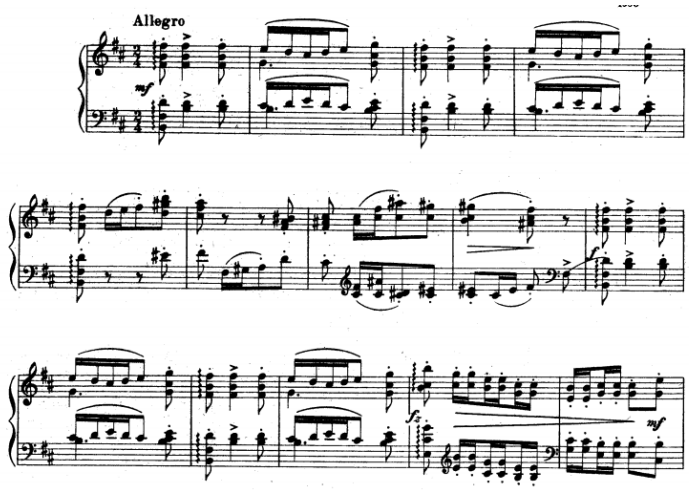


Figure A.17: Hard Piece 2

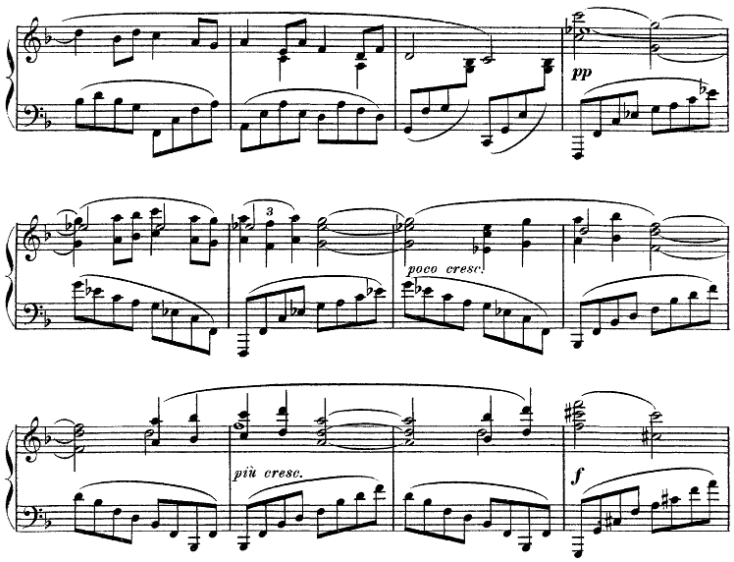


Figure A.18: Hard Piece 3

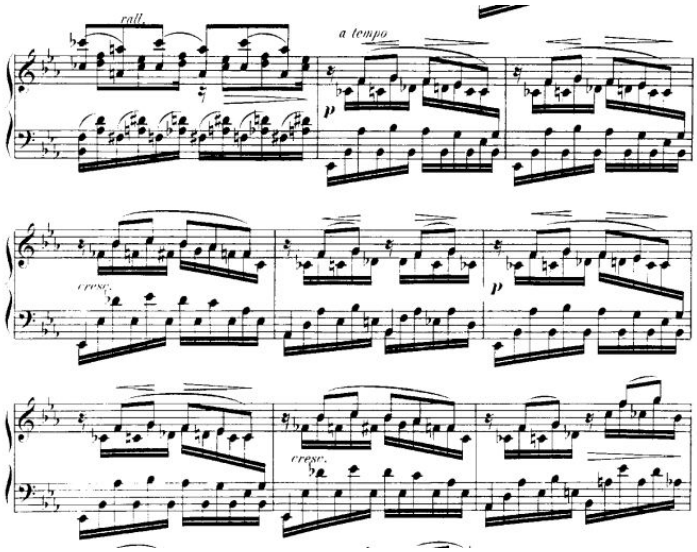


Figure A.19: Hard Piece 4

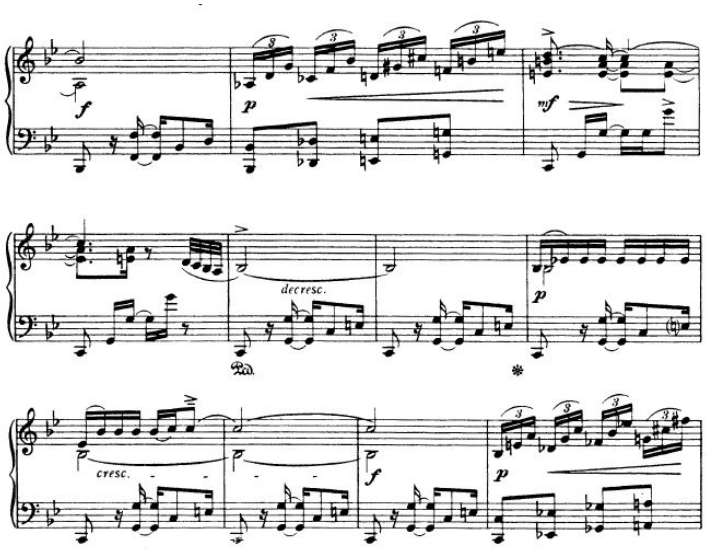


Figure A.20: Hard Piece 5

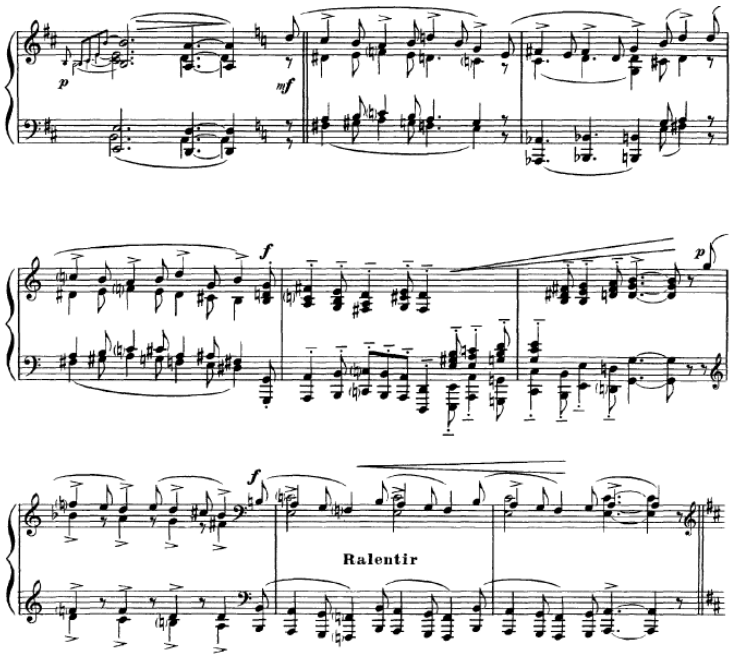


Figure A.21: Hard Piece 6

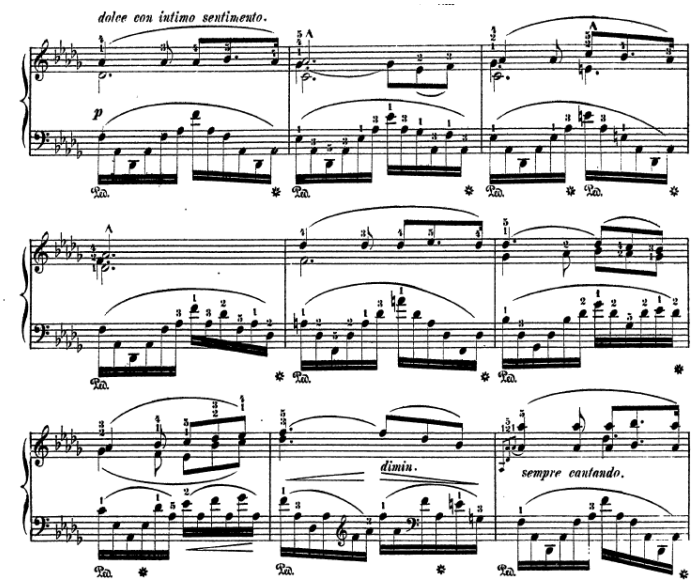


Figure A.22: Hard Piece 7

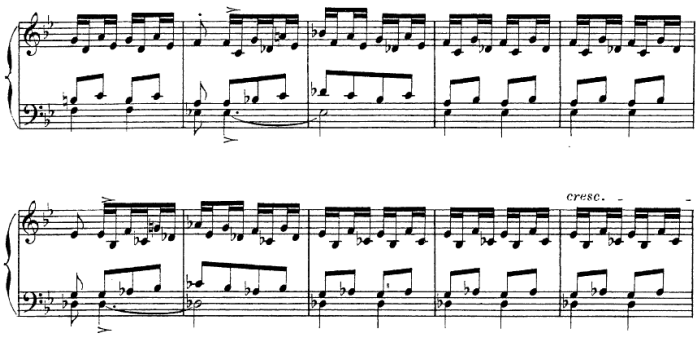


Figure A.23: Hard Piece 8

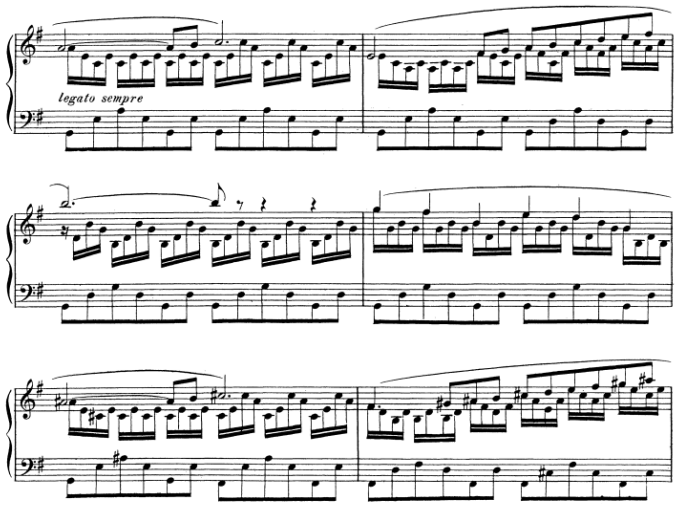


Figure A.24: Hard Piece 9

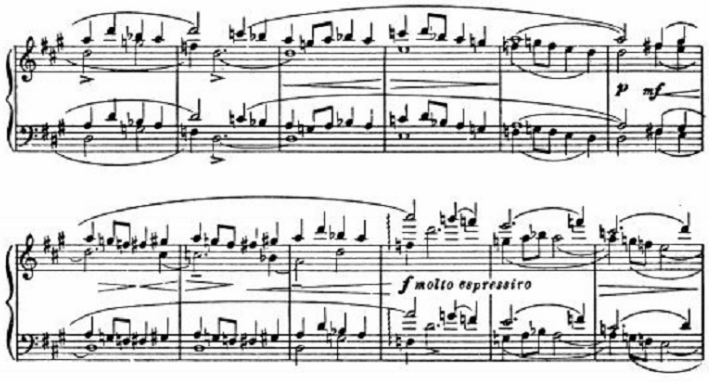


Figure A.25: Hard Piece 10

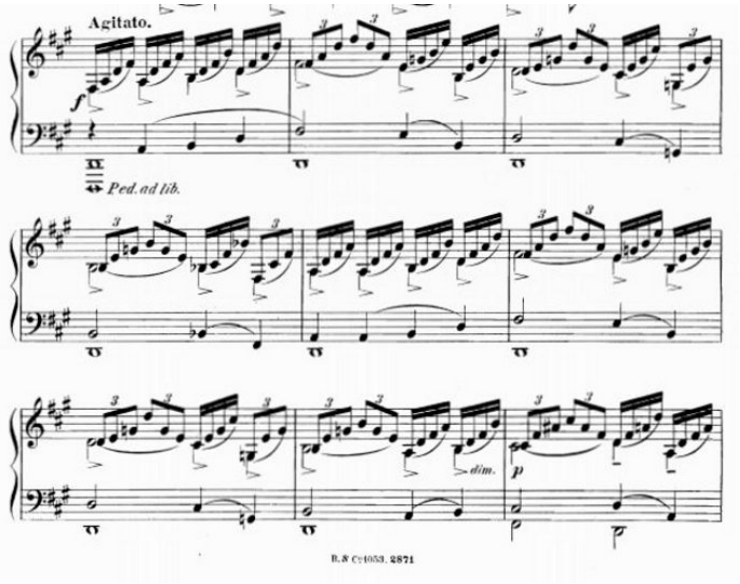


Figure A.26: Hard Piece 11

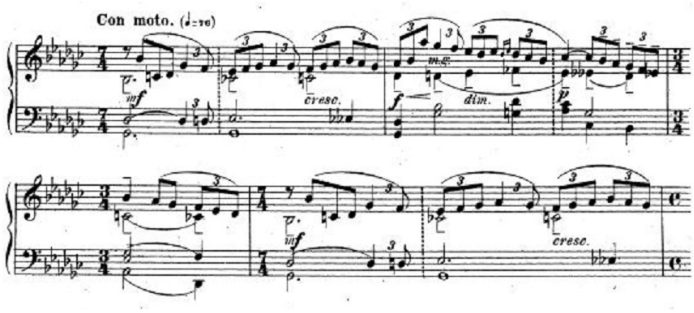


Figure A.27: Hard Piece 12

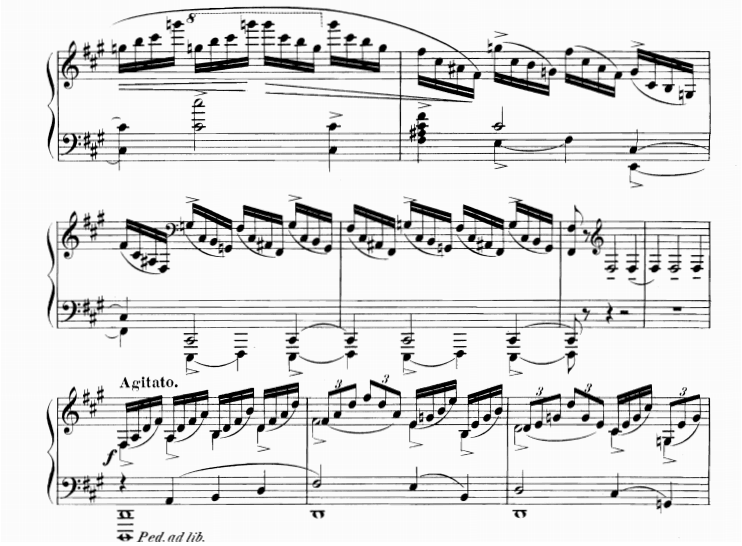


Figure A.28: Hard Piece 13

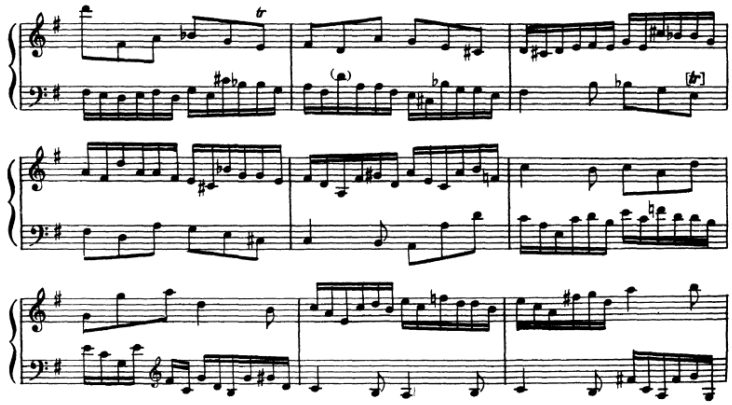


Figure A.29: Hard Piece 14

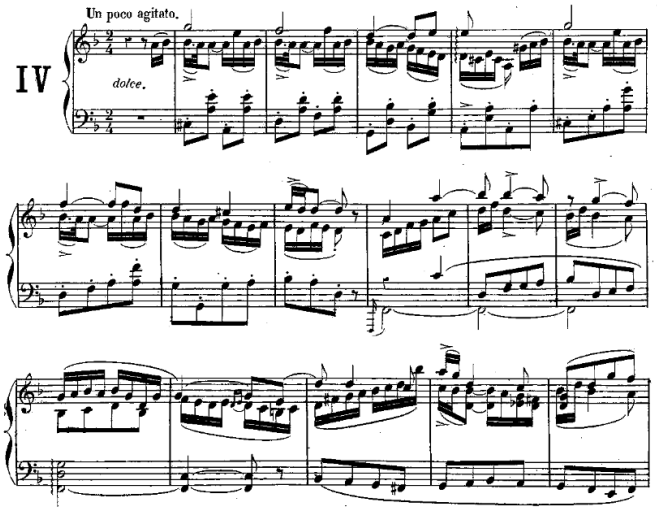


Figure A.30: Hard Piece 15

Appendix B

Questionnaires

FORM OF CONSENT

Please consider this information carefully before deciding whether to participate in this research.

Purpose of research: To understand the blood oxygenation levels of the prefrontal cortex during computer usage.

What you will do in this study: If you decide to participate, you will first be asked to complete a brief questionnaire regarding basic demographic information. Once you are finished, the on-site researcher will fit you with two optical probes placed flush against your forehead and secured using a cotton headband. The probes will be used to measure blood flow to your prefrontal cortex during the activity. You will then be presented with a series of visual stimuli on a computer display, and perform keyboard and mouse actions as prompted and/or interact with the computer using an electronic musical keyboard. Following, you will receive a standard questionnaire to complete regarding the session experience.

Time required: The session will take approximately 60-120 minutes to complete.

Risks and Discomfort: There are no foreseeable risks with this study.

Benefits: There are no direct personal benefits, but you may find it interesting to learn about the topics presented in this session.

Compensation: You will receive $10/hour at the end of the evaluation.

Confidentiality: Your responses will be recorded, and additionally your actions may be videotaped. No identifying features (name, face, etc) will be present in the recorded data. This information will be kept strictly confidential and used only for research purposes. The documentation will be identifiable only by a randomly assigned ID number. All information will be kept in a secure, locked office. When research results are reported, responses will be aggregated and described in summary.

Participation and withdrawal: Your participation is completely voluntary. Should you decide at any time during the study that you no longer wish to participate, you may withdraw your consent and discontinue your participation without penalty or loss of benefits.

To Contact the Researcher: If you have questions or concerns about this research, please contact the principal investigator: Robert Jacob. Human-Computer Interaction Laboratory; 161 College Avenue, Medford MA 02155; email: jacob@cs.tufts.edu phone: 617-627-2225.

Whom to contact about your rights in this research, for questions, concerns, suggestions, or complaints that are not being addressed by the researcher, or research-related harm: Dr. Lara Sloboda at the Office of the Institutional Review Board by email at Lara.Sloboda@tufts.edu or by phone at (617) 627-3417.

Figure B.1: Consent Form for Expts 1, 2, and 3, Page 1

Signature: I confirm that I understand the purpose of the research and the study procedures. I understand that I may ask questions at any time and can withdraw my participation without prejudice. I have read this consent form. My signature below indicates my willingness to participate in this study.

_____

Participant Signature Date

Consent to be videotaped: I confirm that I understand the purpose of the research and the study procedures. I understand that I may ask questions at any time and can withdraw my participation without prejudice. I have read this consent form. My signature below indicates my willingness to be videotaped in this study.

_____

Participant Signature Date

_____

Printed Name of Participant

_____

Researcher Signature Date

_____

Printed Name of Researcher

Figure B.2: Consent Form for Expts 1, 2, and 3, Page 2

How many years of musical training have you received?

Instrument (including voice)	Years	Months
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

Have you majored or minored in Music at college?

Major Minor Neither

Have you studied music theory?

Yes No

If yes, where did you study theory?

Classes at college _____

Private piano lessons_____

Other (please specify)_____

When did you start playing the piano?

_____

Figure B.3: Questionnaire for Expts 2 and 3, Page 1

Do you still play the piano?

Yes No

If yes, how frequently do you play the piano now? Please pick nearest frequency.

Every day _____

Once a week _____

Once a month _____

Once a year _____

Which hand do you feel most comfortable playing the piano with?

Left Right Equally comfortable with both

What kind of music do/did you play on the piano?

Classical _____

Jazz _____

Blues _____

Modern (pop/rock) _____

Other (please specify) _____

How often do/did you improvise on the piano?

Never Rarely Sometimes Often Always

How often do/did you sightread while playing the piano?

Never Rarely Sometimes Often Always

Figure B.4: Questionnaire for Expts 2 and 3, Page 2

How comfortable do you feel with sightreading?

Not at all comfortable Slightly comfortable Moderately comfortable

Very comfortable Completely comfortable

What kind of music do you listen to?

Classical _____

Jazz _____

Blues _____

Pop/rock _____

Electronic _____

Other (please specify) _____

Have you ever used Band in a Box?

Yes No

If you have ever used any other musical software for playing, improvisation or composition, please specify.

_____

_____

_____

_____

_____

Figure B.5: Questionnaire for Expts 2 and 3, Page 3

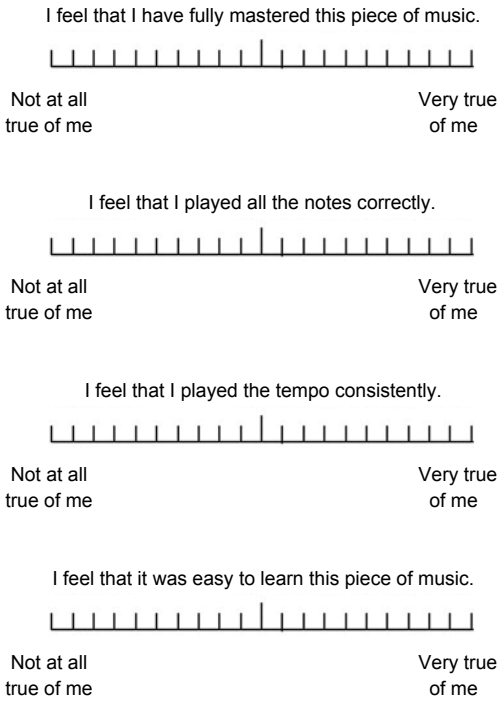


Figure B.6: Questionnaire for Expt 2, Page 4

The Creativity Support Index

Exploration

It was easy for me to explore many different options, ideas, designs, or outcomes without a lot of tedious, repetitive interaction.

Agree |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| Disagree

Collaboration

I was able to work together with others easily while doing this activity.

Agree |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| Disagree

Engagement

I was very absorbed/engaged in this activity - I enjoyed it and would do it again.

Agree |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| Disagree

Effort/Reward Tradeoff

What I was able to produce was worth the effort required to produce it.

Agree |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| Disagree

Tool Transparency

While I was doing the activity, the tool/interface/system "disappeared," and I was able to concentrate on the activity.

Agree |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| Disagree

Expressiveness

I was able to be very expressive and creative while doing the activity.

Agree |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| Disagree

Figure B.7: Questionnaire for Expt 3, Page 4

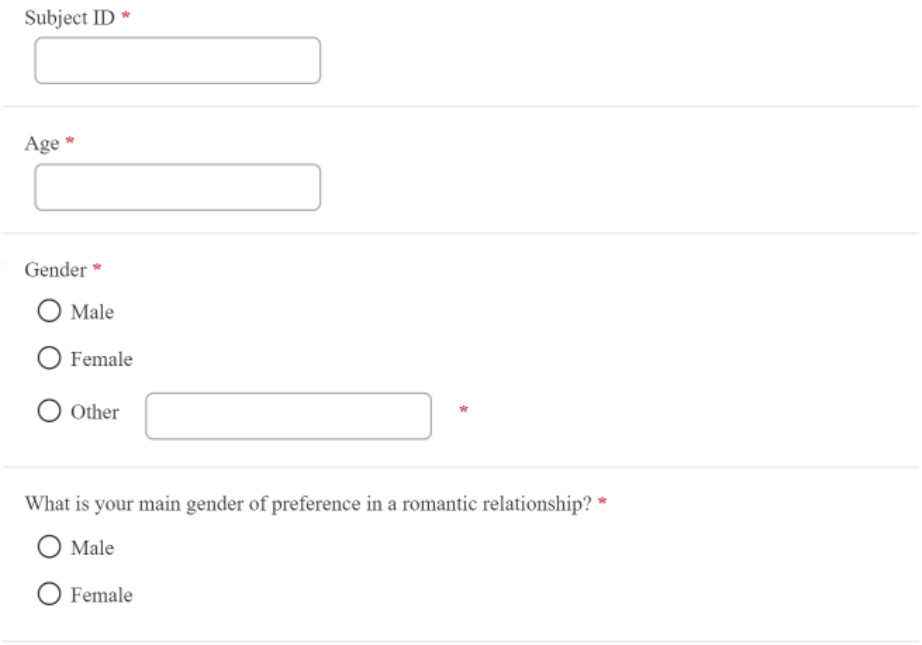


Figure B.8: Questionnaire for Expt 4

How stressed do you feel right now? *

- ○ 1 - Very Unstressed
- ○ 2
- ○ 3
- ○ 4 - Neither Stressed nor Unstressed
- ○ 5
- ○ 6
- ○ 7 - Very Stressed

How attractive did you find the virtual agent's *physical appearance* during this last set of questions? *

- ○ 1 - Very Unattractive
- ○ 2
- ○ 3
- ○ 4 - Neither Attractive nor Unattractive
- ○ 5
- ○ 6
- ○ 7 - Very Attractive

How attractive did you find the virtual agent's *voice* during this last set of questions? *

- ○ 1 - Very Unattractive
- ○ 2
- ○ 3
- ○ 4 - Neither Attractive nor Unattractive
- ○ 5
- ○ 6
- ○ 7 - Very Attractive

Figure B.9: Questionnaire for Expt 4

How *accurate* did you find the virtual agent during this last set of questions? *

- ○ 1 - Very Inaccurate
- ○ 2
- ○ 3
- ○ 4 - Neither Accurate nor Inaccurate
- ○ 5
- ○ 6
- ○ 7 - Very Accurate

How much did you *trust* the virtual agent during this last set of questions? *

- ○ 1 - Strongly Mistrusted
- ○ 2
- ○ 3
- ○ 4 - Neither Trusted nor Mistrusted
- ○ 5
- ○ 6
- ○ 7 - Strongly trusted

How much did you *like* the virtual agent during this last set of questions? *

- ○ 1 - Strongly Disliked
- ○ 2
- ○ 3
- ○ 4 - Neither Liked nor Disliked
- ○ 5
- ○ 6
- ○ 7 - Strongly Liked

Figure B.10: Questionnaire for Expt 4