MODELING USER EMOTION IN INTERACTIVE PLAY ENVIRONMENTS: A FUZZY PHYSIOLOGICAL APPROACH

by

Regan Lee Mandryk

B.Sc. University of Winnipeg 1997

M.Sc. Simon Fraser University 2000

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in the School
of
Computing Science

SIMON FRASER UNIVERSITY

FALL 2005

APPROVAL

NAME:	Regan L. Mandryk
DEGREE:	Doctor of Philosophy
TITLE OF THESIS:	Modeling User Emotion in Interactive Play Environments: A Fuzzy Physiological Approach
EXAMINING COMMITTEE:	Torsten Möller, chair

Dr. Kori M. Inkpen, Senior Supervisor
Associate Professor, Faculty of Computer Science
Dalhousie University

Dr. Thomas W. Calvert, Senior Supervisor
Professor, School of Computing Science
Simon Fraser University

Dr. M. Stella Atkins, Supervisor
Professor, School of Computing Science
Simon Fraser University

Dr. Lyn Bartram, Supervisor
Assistant Professor, School of Interactive Arts and Technology
Simon Fraser University, Surrey Campus

Dr. Kellogg S. Booth, Supervisor
Professor, Department of Computer Science
University of British Columbia

Dr. John Dill, SFU Examiner
Professor, School of Engineering Science
Simon Fraser University

Dr. Scott Hudson, External Examiner
Professor, Human-Computer Interaction Institute
School of Computer Science, Carnegie Mellon University

DATE APPROVED: Dec. 13/05

STATEMENT OF ETHICS APPROVAL

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

(a) Human research ethics approval from the Simon Fraser University Office of Research Ethics,

or

(b) Advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University;

or has conducted the research

(c) as a co-investigator, in a research project approved in advance,

or

(d) as a member of a course approved in advance for minimal risk human research, by the Office of Research Ethics.

A copy of the approval letter has been filed at the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, BC, Canada

ABSTRACT

Researchers are integrating emerging technologies into interactive play environments, and established game markets continue to expand, yet evaluating play environments is challenging. While task performance metrics are commonly used to objectively and quantitatively analyse productivity systems; with play systems, the *quality of the experience*, not the *performance of the participant* is important. This research presents three experiments that examine users' physiological signals to continuously model user emotion during interaction with play technologies. Modeled emotions are powerful because they capture usability and playability, account for user emotion, are quantitative and objective, and can be represented continuously.

In Experiment One we explored how physiological signals respond to interaction with play technologies. We collected a variety of physiological measures while observing participants playing a computer game in four difficulty conditions, providing a basis for experimental exploration of this domain.

In Experiment Two we investigated how physiological signals differ between play conditions, and how physiological signals co-vary with subjective reports. A different physiological response was observed when playing a computer game against a co-located friend versus a computer. When normalized, the physiological results mirrored subjective reports.

In Experiment Three we developed a method for modeling emotion using physiological data. A fuzzy logic model transformed four physiological signals into arousal and valence. A second fuzzy logic model transformed arousal and valence into five emotions: boredom, challenge, excitement, frustration, and fun. The modeled emotions' means were evaluated with test data, and exhibited the same trends as the reported emotions for fun, boredom, and excitement, but modeled emotions revealed differences between three play conditions, while differences between reported emotions were not significant.

Mean emotion modeled from physiological data fills a knowledge gap for objective and quantitative evaluation of entertainment technologies. Using our technique, user emotion can be analyzed over an entire experience, revealing variance within and between conditions. This continuous representation has a high evaluative bandwidth, and is important because the *process*, not the *outcome* of playing determines success. The continuous representation of modeled emotion is a powerful evaluative tool, that when combined with other approaches, forms a robust method for evaluating user interaction with play technologies.

Keywords:

User Interfaces, human-computer interaction, emotion, play, computer games, fun, evaluation methodology, physiology, GSR, EMG, HR, fuzzy logic, affective computing

DEDICATION

For Kevin

For pushing me, pulling me, and
walking beside me

For supporting my hockey habit
and keeping me supplied with
pink pens and black sweats

I am fortunate beyond words to
have you as my partner and
friend

ACKNOWLEDGEMENTS

There are so many people to thank for helping me along this journey. I can't possibly name all of the people that supported me, but there are some special thanks that need to go out to those who have profoundly impacted me and my research.

For generous financial support, I would like to thank NSERC, NECTAR, NewMIC, and EA Canada. My experiments would not have been possible without the assistance of many people. Thanks to Thecla Schiphorst at SFU Surrey, John Buchanan at EA, the GrUVi Lab at SFU, the Imager Lab at UBC, and the NRC for lending equipment and providing the participant incentives.

Many groups have provided me with a home during the course of my research. For making me welcome, providing me with desk space, and bunging up my inbox with departmental e-mail, I would like to thank the GrUVi Lab at SFU, the Affectionate Computing Group at SFU Surrey, NewMIC, the Imager Lab at UBC, and the EDGE Lab at Dalhousie University. There are far too many people that I've met at each of these institutions to thank them all personally, but I would like to mention Dr. Torsten Möller, and Dr. Ted Kirkpatrick at SFU; Sang Mah, Gordon Pritchard, Glenn Davies, Alan Boykiw, Dr. David Darvill, Julia Rylands, Todd Zimmerman, Brian Corrie, and Dr. Rodger Lea from my time at NewMIC and SFU Surrey; and Dr. Joanna McGrenere and Dr. Jason Harrison from UBC. In addition, I've been fortunate to interact with numerous students from each of the research institutions that opened

their doors to me. Thanks to all of the students that I met along the way, and specifically to Diego Maranan for his work on the False Prophets project, which defined my research direction. The members of the EDGE Lab, both past and present, provided friendship, advice, and encouragement during some painful all-nighters. Thanks also to all of the students from NewMIC and the IDRG at UBC. Everyone who commits to a Ph.D. should have a colleague with whom to take the journey. I am deeply fortunate to have had Dr. Stacey Scott, a fellow student and friend who I respect and admire, who went through every step with me, and supported me when I had 'the academic bends'.

A special thanks has to go to Hiromi Matsui at SFU for providing endless support, great ideas, and the 'Hiromi Express'. The CS administrative staff, the Network Support Group and Technical Support Staff in CS at SFU also provided a lot of assistance throughout my time there. A special thanks goes to sumo Kindersley for always giving me answers, friendship, and Diet Pepsi.

My committee has provided me with guidance, feedback and support. First, thanks to Dr. Kori Inkpen, my advisor and friend. I deeply appreciate her advice and supervision, whether it occurred during an office meeting, over beer at the bar, or during a 2am powwow complete with low-fat snacks. She inspired my original interest in collaborative play, and supported my research direction, even where it diverged from her own, and her generosity and kindness has made the Mary Ellen Carter nothing more than a snappy song about a boat. Dr. Tom Calvert has so much wisdom and experience that I am truly grateful for his advice and am fortunate for his

imprint on this dissertation. Dr. M. Stella Atkins provided a perspective to my research that was invaluable, especially in the later stages that involved the use of fuzzy logic for modeling emotion. Thanks also to Dr. Kellogg Booth and Dr. Lyn Bartram for their guidance and revisions, and thanks to my examiners, Dr. Scott Hudson and Dr. John Dill for their valuable comments.

Thanks to my parents, Don and Leona, my sister Kara and brother Jason, and the rest of my family who supported me, inspired me, and taught me to stand tall and reach beyond my limits. My parents taught me to be curious, and instilled in me the desire for lifelong learning. Thanks to Kevin, who encouraged me, put up with me, and kept me from going insane. He sat through endless brainstorming conversations with me when I'm sure he would rather have been doing something else. I feel lucky to have a partner who understands both my work and that I don't do dishes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 INTRODUCTION

1.1 Background

Computer games have grown during recent years into a popular entertainment form with a wide variety of game types and a large consumer group spread across the world. An increasing number of people are playing electronic games, placing them among other favorite leisure activities, like reading books and watching films [55]. When surveyed on the most fun entertainment activities in the year 2000, 35 percent of all Americans identified computer and video games, whereas watching television fell second at 18 percent, followed by surfing the internet (15%), reading books (13%) and going to the movies (11%) [55].

On-line gaming has offered people new means of having social interaction with gamers in other locations, and has let gamers access and play out fantasy-driven identities that they are unable to manifest in the real world [132]. In 2002, the percentage of gamers that play online rose to 31 percent up from 24 percent the year before [55]. In 2004, the percentage increased to 43% [55]. Within the games, or through the use of websites based around the games, dedicated communities have formed which have created new content, sometimes leading to commercial ventures. Electronic game play, however, is not limited to home use. Game parlors and

LAN (Local Area Network) parties are becoming a popular means to play online games [49]. Also, 37 percent of Americans who own game consoles (or computers used to play games), report that they also play games on mobile devices such as PDAs (Personal Digital Assistants) and mobile phones [55]. The popularity of computer games has made them important carriers of culture and trends but also a vehicle for the development and deployment of new hardware, software and user interface techniques.

According to the Interactive Digital Software Association (www.idsa.com), revenue from the computer and video game software industry in the US nearly doubled from $3.2 billion in 1994 to $6.35 billon in sales from 225 million units in 2001 (up 7.9% and 4.5% from 2000 respectively). In the same year in the United States, movie box office grosses were $8.41 billion [55]. In 2004, game sales increased to $7.3 billion, inching closer to the total revenues from the film industry. In Great Britain in the year 2000, the entertainment software industries grossed £300 million more than the British cinema box offices and almost double that of home video rentals [32].

1.2 Motivation

In addition to growth in traditional computer and console games, emerging technologies in ubiquitous computing and ambient intelligence offer exciting new interface opportunities for play technology, as evidenced in a recent growth in the number of conference workshops and research articles devoted to this topic [6, 8, 9, 71, 72]. Our research team is interested in employing these new technologies to foster interactions between users in co-located, collaborative entertainment environments.

We want technology not only to enable fun, compelling experiences, but also to enhance interaction and communication between players.

We have created a few novel game environments with the goal of enhancing interaction between players and to create a compelling experience [22, 75, 76]. One of our game environments, False Prophets [76], was a hybrid board/video game, played on an interactive tabletop with a tangible interface. After creating False Prophets, which is described in more detail in Appendix 2, we wanted to determine whether our novel game features created an interactive and engaging experience. However, we found that none of the current evaluation methodologies were robust enough to answer our research questions. Other researchers have also used emerging technologies to create entertainment environments [6, 8, 51, 71, 72], yet evaluating the success of these new interaction techniques and environments is an open research challenge for the ubiquitous gaming community. Upon further examination, we noted that traditional computer game developers were also suffering from a lack of effective evaluation methods.

Traditionally, human-computer interaction research (HCI) has been rooted in the cognitive sciences of psychology and human factors, and in the applied sciences of engineering, and computer science [94]. Although the study of human cognition has made significant progress in the last decade, the notion of emotion is equally important to design [94], especially when the primary goals are to challenge and entertain the user. This emotion-centric approach presents a shift in focus from usability analysis to human experience analysis.

The first issue prohibiting good evaluation of entertainment technologies is the inability to define what makes a system successful. We are not interested in traditional performance measures, we are interested in what kind of emotional experience is provided by the play technology and environment [96]. Although traditional usability measures may still be relevant, they are subordinate to the emotional experiences resulting from interaction with play technologies and with other players in the environment.

Once we determine what makes an entertainment system successful, we need to resolve how to measure the chosen variables. Unlike performance metrics, the measures of success for collaborative entertainment technologies are more elusive. The current research problem lies in what emotions to measure, and how to measure them. These metrics will likely be interesting to researchers and developers of games and game environments.

1.3 Overview of Research

Our goal is to develop an evaluation methodology for entertainment environments that:

1. captures usability and playability through metrics relevant to ludic[1] experience;
2. accounts for user emotion;
3. is objective and quantitative; and
4. has a high evaluative bandwidth.

[1] Of, or referring to play or playfulness [26].

Researchers in human factors have used physiological measures as indicators of mental effort and stress [137]. Psychologists use physiological measures to differentiate human emotions such as anger, grief, and sadness [31]. However, physiological data have not been employed to identify a user's emotional states such as fun and excitement when engaged with entertainment technologies. Based on previous research on the use of psychophysiological techniques, we believe that capturing, measuring, and analyzing autonomic nervous system (ANS) activity will provide researchers and developers of technological systems with access to the emotional experience of the user. Used in concert with other evaluation methods (e.g. subject reports and video analysis), a complex, detailed account of both conscious and subconscious user experience could be formed.

This dissertation describes a research program designed to test the efficacy of physiological measures for use in evaluating player experience with collaborative entertainment technologies. We have three main conjectures:

Conjecture A: *Physiological measures can be used to objectively measure a player's experience with entertainment technology.*

Conjecture B: *Normalized physiological measures of experience with entertainment technology will correspond to subjective reports.*

Conjecture C: *Physiological metrics can be used to model user emotional experience when playing a game, providing continuous, quantitative, and objective metrics of evaluation for interactive play technologies.*

1.3.1 Organizational Overview

We begin by describing current techniques for evaluating interactive technologies in Chapter 2. Many of these methods were developed to evaluate productivity applications and environments. As such, we discuss how the methods have been adapted to evaluate play technologies. Although some evaluation methods have been successfully used to evaluate game and play environments, Section 2.3.4 describes where current methods fall short, and shows the lack of objective and quantitative methods for evaluating play.

In Chapter 3 we introduce some of the psychological concepts relevant to our research. These concepts include an overview of the Human Nervous System (section 3.1), arousal (section 3.2), valence (section 3.3), habituation (section 3.4), flow (section 3.5), and emotion (section 3.6).

To provide an introduction for readers unfamiliar with physiological measures, in Chapter 4 we briefly introduce the physiological measures used in our research, describe how these measures are collected, and explain their inferred meaning. Metrics relating to electrodermal activity (section 4.1), cardiovascular activity (section 4.2), respiration (section 4.3), and muscle activity (section 4.4) are discussed. Chapter 4 also presents information on indexing psychological events from physiological data

(section 4.5). This includes competing theories on the classification of emotion (section 4.5.1), and issues and limitations associated with sensing physiological responses (section 4.5.2).

In the field of Human Factors, researchers have been using physiological signals as an evaluation metric in many domains. As such, Chapter 5 provides related literature on using physiological signals as metrics of evaluation in laboratory tasks (section 5.1), in field tasks (section 5.2), in dispatch, air traffic control and simulator tasks (section 5.3), and with adaptive technologies (section 5.4). Although there has been little research on using physiological signals as evaluation metrics for interaction with computer technologies, work in the domain of HCI is discussed in section 5.5.

In Chapter 5 we also introduce the research area of affective computing (section 5.6), which is computing that relates to, arises from, or deliberately influences emotion [101]. Physiological signals have been used as input to interactive systems, and although we propose to use physiological signals as an evaluation methodology, in section 5.6.1 we present relevant research on using body signals as input. The chapters on related literature close with a brief examination of wearable biometric sensors in section 5.6.2.

The remainder of the dissertation presents research designed to investigate the applicability of physiological measures as indicators of human experience with entertainment technologies. We describe three experiments that we designed to test our main conjectures. Throughout the experiments, we record users' physiological, verbal and facial reactions while they play NHL 2003 by EA Sports in different play

conditions. We apply post-processing techniques to correlate an individual's physiological data with their subjective reported experience.

In Experiment One, we manipulated game difficulty and explored how a user's physiological signals responded to interaction with play technologies. Chapter 6 describes the participants (section 6.1), play conditions (section 6.2), experimental setting (section 6.3), and data analysis techniques (section 6.4). Results of the experiment are presented in section 6.5. We experienced some methodological problems in Experiment One, which are discussed in section 6.6. The issues that we experienced, and the results of the experiment allowed us to generate rules for conducting experiments in this domain, which are presented in Chapter 6.

Based on the lessons we learned, and the results from Experiment One, we conducted Experiment Two. Experiment Two investigated how physiological signals co-varied with subjective reports, lending support for Conjecture A, that *physiological measures can be used to objectively measure a player's experience with entertainment technology,* and Conjecture B, that *normalized physiological measures of experience with entertainment technology will correspond to subjective reports.* We manipulated game opponent (co-located friend or stranger), and Chapter 7 describes the participants (section 7.1), play conditions (section 7.2), experimental setting (section 7.3), and data analysis techniques (section 7.4). Results of the experiment are presented in section 7.5. Because of the methodological issues that we experienced in the first experiment, and the subsequent rules that we developed, we made many adjustments in Experiment Two to our data collection techniques and experimental

design. Although these changes are described in the relevant aforementioned sections, we also highlight the adjustments, and their impact in section 7.6.

Based on the knowledge acquired in Experiments One and Two, in Experiment Three we developed a method for modeling emotion, using physiological signals. Due to the success of Experiment Two, we collected data in three play conditions: against a co-located friend, against a co-located stranger, and against the computer. We developed a fuzzy logic model that transformed four physiological signals into values of arousal and valence. A second fuzzy logic model transformed the arousal and valence values into continuous values for five emotions: boredom, challenge, excitement, frustration, and fun.

Chapter 8 presents details on how we collected the data and generated the modeled emotions. In section 8.2, we present a brief introduction to fuzzy logic, then present the details of how we modeled arousal-valence space in section 8.3. Our modeled arousal and valence values compared favorably to values generated using a brute force approach (section 8.3.5). There were some outstanding issues related to modeling arousal and valence, which are discussed in section 8.3.6. Our second model, which transforms arousal and valence into the five modeled emotions is presented in section 8.4. The outstanding issues with modeling emotion are also discussed in this section.

Chapter 9 presents how we used the model to objectively and quantitatively evaluate emotional experience during interaction with NHL2003 by EA Sports. Results for modeled emotions (section 9.1) are presented along with results for reported emotions (section 9.2) for the same five emotions. The modeled emotions were successfully

compared to subjective reports in section 9.3, supporting Conjecture C, that *physiological metrics can be used to model user emotional experience when playing a game, providing continuous, quantitative, and objective metrics of evaluation for interactive play technologies.* Although successful, our modeled emotions suffer from scaling issues, which are presented in section 9.4 along with potential solutions. In addition to providing a quantitative and objective methodology for evaluating user interaction with play technologies, modeled emotions can be represented continuously, yielding a method with a very high evaluative bandwidth. The continuous nature of modeled emotions is highlighted in section 9.5.

Finally, we conclude the dissertation with a summary of the results and contributions in section 10.1, and discuss our plans for future work in section 10.3. Chapter 11 provides a list of the references used in this dissertation.

Twelve appendices are included at the end of the dissertation. Appendix 1 lists the abbreviations and acronyms used throughout the dissertation. Appendix 2 gives more detailed information on False Prophets, the game environment that we developed which motivated this research direction. Appendix 3 provides results from electrode placement tests that we conducted to ensure valid results from our sensor placement. Appendix 4 contains the consent form used in all of our experiments as required by the guidelines for conducting research on human participants from the Simon Fraser University Research Ethics Board. Appendices 5 through 11 include questionnaires used in the three experiments, while Appendix 12 through Appendix 14 contain extra information and results from the process of modeling emotion.

1.4 Summary

Researchers are using emerging technologies to develop novel play environments, while established computer and console game markets continue to grow rapidly. Even so, evaluating the success of interactive play environments is still an open research challenge. Both subjective and objective techniques fall short due to limited evaluative bandwidth, and there remains no corollary in play environments to task performance with productivity systems. In addition, we want to incorporate a user's attitudes, behaviours, and emotions into an evaluation.

This dissertation presents an investigation into the efficacy of a user's physiological signals as evaluators of interaction with play technologies. This approach could be powerful as it captures usability and playability through metrics relevant to ludic experience, accounts for user emotion, is quantitative and objective, and is represented continuously over a session.

Chapter 2 CURRENT METHODS OF EVALUATING ENTERTAINMENT TECHNOLOGIES

Methods of evaluating computing technologies range from rigorous to casual, and can be qualitative or quantitative, subjective or objective, or some hybrid approach. The common methods of evaluating user interaction with technology are described in this section. We include descriptions of subjective and objective techniques for hypothesis testing and evaluation. We do not present methods of hypothesis generation common to some social sciences (e.g., ethnography) as these methods are used for forming theories used to inform the design of technology rather than to evaluate technology in any stage of development. Although these social techniques could be utilized to study gamers and gaming culture, we are more interested in how to evaluate specific technologies.

2.1 Subjective Methods of Evaluation

Subjective measures of evaluation for human computer interaction (HCI) typically include questionnaires, interviews, and focus groups. There are other subjective evaluation techniques commonly used in the social sciences such as ethnography and social observational schemes, and recently some of these social science techniques have been adapted for use in HCI research. For example, contextual inquiry [52] and rapid ethnography [84] have been used to discover trends in work practices for

technology-rich office environments. These subjective social science methodologies are generally used for *hypothesis generation,* using qualitative techniques, rather than *hypothesis testing,* using quantitative techniques [35, 79, 81]. As such, they are not discussed in detail in this section. The subjective evaluation methods of questionnaires, interviews, and focus groups are presented.

2.1.1 Questionnaires

Techniques such as questionnaires and surveys that require users to rate their experience through a series of statements and questions are common and straightforward methods of subjectively evaluating technologies [79, 122]. Questionnaires and surveys are considered to be generalizable, convenient, amenable to rapid statistical analysis, and easy to administer. The large amount of data that can be gathered from surveys offers the results a sense of conviction [122]. Rating a statement or a user interface feature using a series of bipolar semantically anchored items or a Likert scale provides numerical data that can be analyzed using non-parametric statistical methods.

Some drawbacks of using questionnaires or surveys are that: survey techniques aren't conducive to finding complex patterns; questionnaires can invade privacy; and because subjective reports are cognitively mediated, they may not correspond to the actual experience of the survey participant [79, 149]. Knowing that their answers are being recorded, participants will sometimes answer what they think you want to hear, perhaps without even realizing it.

2.1.2 Interviews and Focus Groups

Interviews are different from questionnaires in that they cannot be administered by a computer or on paper, but involve an experimenter asking the questions and recording the answers given by the participant [92]. Thus, interviews can be more free-form than questionnaires since the experimenter can rephrase difficult questions and prompt participants for more depth on any given question [92, 122]. This makes interviews harder to analyze quantitatively since not all subjects may be asked the same questions under the same conditions. Conversely, they often provide rich descriptions and may elucidate the quantitative results from questionnaires. Interviewers must be careful to not bias the participant's responses and to ask questions in a neutral, non-leading manner.

Focus groups are a fairly informal technique that involves bringing a small number of participants together with a moderator to discuss user needs and feelings [92]. A focus group should be free-flowing from a participant's perspective, but the moderator should maintain the focus. Focus groups are sometimes preferred over interviews due to the time saved by interviewing multiple people simultaneously, but also because of the spontaneous reactions and ideas that emerge through the participants' interactions [92]. Focus groups are also limiting in that the results are always qualitative and subjective. In addition, participants' opinions may be swayed by other, more vocal participants in the group [92].

Focus groups are perhaps the most utilized method for evaluating games [41]. Game companies use surveys less often due to the expertise needed in analyzing the data

[41], but surveys can yield results not available from focus groups (e.g., quantitative evaluation). Subjective techniques have been used to evaluate game usability as well as game playability, and the advantage of subjective techniques is that a researcher or developer can access information related to user preferences and attitudes, an important factor in playability evaluation.

2.2 Objective Methods of Evaluation

One of the most common methods of objectively evaluating interactive technologies is using task performance, but other objective measures gathered through video analyses can be equally informative. In this section, we present an overview of some objective methods of evaluation and the techniques used to obtain them.

2.2.1 Task Performance

Depending on the task, a number of task performance indicators can be used. There are comprehensive general lists of task performance measures that can be adapted for most experimental situations [92, 122, 126]. In addition, certain specific research areas have well-studied and well-documented methods of evaluating task performance. For example, text-entry and target selection on devices has been studied from the earliest interactive computer through to recent mobile devices. Fitts presented models for serial tapping tasks in 1954 [36], and discrete aiming tasks in 1964 [37] that have been adapted for use with modern computer interfaces, and are widely used today to analyze and predict movement times for targeting tasks on personal computers, cell phones, and PDAs. Card, Moran, and Newell [15] discussed

a keystroke-level model for user performance in a 1980 issue of *Communications of the ACM*, and this area of research is still being advanced and iteratively evaluated.

General measures of task performance include but are not limited to [92, 126]:

- Task completion time
- The number of user errors
- Percent tasks completed
- The number of system features that can be remembered during a debriefing
- Time spent using the Help functions
- How frequently the Help system solved the user's problem.

After logging this information, standard quantitative statistical methods are used to analyze the data.

Usability testing (in terms of user experience, rather than Quality Assurance), has not been a standard method for evaluating games, although testing techniques offer potential for gathering information related to usability of the interface, as opposed to playability [18, 41]. The time it takes to test games is negligible when considering the time needed to fix the problems that usability testing might find. There has been a recent effort into rapid usability testing, which would more likely be adopted by game companies. Rapid Iterative Testing and Evaluation (the RITE method) was used to evaluate a popular game (Age of Empires II) at Microsoft Games Studios [82]. Although the case study suggests that RITE was useful in evaluating Age of Empires II, RITE has not been tested for general use over a wide variety of games.

In 1982, Malone [73] published one of the earliest papers on game evaluation. The goal was to bring the appealing elements of games into productivity applications, to make tedious productivity software more enjoyable to use. He created eight versions of an educational math game, successively removing motivational features from each version. Participants were given the option of playing their version of the game or an entirely different control game. The primary measure of appeal for each of the versions was how long the participants played their version of the game, as compared to a control game. This measure of appeal correlated with subjective reports of how well the participants liked the game. Although this worked well in the early 1980s, games are so complex now, with so many motivating attributes, that it would be difficult to separate out game features in order to test their appeal. However, measuring the appeal of a game by how long participants choose to play is still a valid, but limited approach.

2.2.2 Observational Techniques and Video Analysis

Observational data recorded on video and in computer-generated logs may include data about the system (e.g., modes and outputs), the environment (e.g., interruptions, network load), or about the user's behaviour (e.g., eye movements, gestures, verbalizations, facial expressions, etc.) [35]. Analysis techniques of observational data from video include conversation analysis, verbal and non-verbal protocol analysis, cognitive task analysis, and discourse analysis [35].

Sanderson and Fisher have described Exploratory Sequential Data Analysis (ESDA) techniques, which are empirical ways of seeking answers to research questions

through the use of observational data under the guidance of formal concepts [35, 117]. In addition, ESDA techniques encompass the three broad traditions of observational research - *behavioural*, *cognitive*, and *social* traditions. In the behavioural tradition, researchers usually construct questions that can be answered objectively and quantitatively, using the scientific method [35, 138]. Analysis tends to focus on events that can be compartmentalized and coded without much interpretation and subjectivity from the researcher. Results are generally quantitative, and stress replicability and generalizability [35, 117]. In the cognitive tradition, verbalizations are as important to analysts as the behavioural data, since verbalizations can offer insights into the cognitive processes underlying and inspiring user action [35, 111]. In the social tradition, questions often focus on the social, interpersonal, cultural or communicative events[2] [35, 40]. Encoding and analysis is an iterative process, grounded in the data itself [81], and the results tend to be qualitative, validated using formal methods of qualitative analysis.

Analyzing video by coding gestures, body language, verbal comments and other subject data as an indicator of human experience is a lengthy and rigorous process that needs to be undertaken with great care [79]. Researchers have to be careful to acknowledge their biases, address inter-rater reliability, and not read inferences where none are present [79]. There is an enormous time commitment associated with observational analysis. The analysis time to data sequence time ratio (AT:ST)

[2] Note that although observation techniques are generally considered objective methods of evaluation, the social tradition focuses on communication, process, and cultural events. These measures are subject to the experimenter's biases and pre-conceived notions, thus are not considered objective measures.

typically ranges from 5:1 to 100:1 [35]. Even five hours of analysis for every hour of data may be too high, so some usability professionals have decreased the analysis time to two hours of analysis for every hour of data, which could result in jeopardizing the quality of the analysis. On the other hand, some cognitive scientists have increased the analysis time to 1000 hours of analysis for every hour of data, for a thorough treatment of the data [35]. As a result of the time commitment, many researchers rely on subjective data for user preference, rather than objective observational analysis.

Given the tremendous time commitment and the need for specialized training, video observational methods have not been widely adopted for the evaluation of games. Since most game development companies do not have the necessary expertise in evaluation, companies like XeoDesign [151] specialize in observational evaluation of game playability and game usability. Using video recordings of what players say and do, questionnaire responses, and verbal and non-verbal emotional cues, expert evaluators assess a player's experience and provide qualitative feedback to clients on how to make their games more fun [65].

2.2.2.1 Facial Expression Analysis

Observational data from video is not limited to verbalizations or observable behaviour. Facial expressions are another commonly observed data source, since facial expressions can be used to identify emotions. A standard method of interpreting facial expressions is to record and analyze them in context. It is quite common to observe a look of concentration, frustration or celebration when people interact with

technology. These "looks" are often associated with body movements and verbal comments.

The study of facial expressions has been centered on the question of whether people use their face to represent emotion. Keltner and Ekman [61] provide a summary of the research domain and the issues and findings that have been encountered since the late 1800s. There have been several significant results relevant to determining whether facial expression can play a role in analyzing user reaction to technology.

Firstly, we know that facial expressions have links to emotions. We smile when we are happy or pleased, and frown when we are discouraged or upset. Our eyes crinkle in myriad ways when we feel different emotions, or are trying to convey our feelings to another person. The unique combination of how each facial element is changed can convey specific emotions. In addition, facial expressions have physiological ties. Our body responds differently when we generate different facial expressions, even when the underlying emotion is not present. For example, expressions of anger, fear and sadness produce greater heart acceleration than other emotions and the expression for anger produces greater finger temperature than that for fear. Although we can't see most of the physiological changes that accompany the making of facial expressions, people are very good at accurately judging facial expressions. In fact, we can judge facial expression of emotion with level of accuracy that exceeds chance (60 to 80% success when chance is calculated between 17 and 50%) [61].

Although facial expressions can be recognized at a rate greater than chance, and a system for coding expressions into distinct categories has been developed [30], this

area of research is still fraught with unanswered questions and methodological issues [61].

2.3 Standard Usability Methods

Much of the recent research on HCI analysis techniques is related to usability analysis. The goal of usability analysis is to inform the design of software and hardware products to ensure that the products adhere to established usability principles, as well as to users' expectations of how the technology will behave. There are some usability methods that do not require the involvement of users; however, many methods involve watching a user work through a set of tasks.

Although some techniques may not be useful for empirical research, these techniques can be adapted or integrated into an experiment, enhancing the empirical data.

2.3.1 Think Aloud Protocols

Asking participants in an experiment to verbalize their experiences is known as a think-aloud protocol. Thinking aloud is a valuable method, used to understand how participants view the technology, and feel about their interactions with the technology [92]. Although this technique is based in psychological research [33, 77], it has been adopted by computer scientists and software developers [90]. Traditionally, there is a significant amount of analysis conducted on the verbal data including verbal transcriptions and coding the utterances according to an iteratively-defined scheme. This process has a high cost in terms of time commitment, requiring an AT:ST ratio of about 25:1 [77, 90]. Other researchers have adopted a "discount" approach to the

think-aloud protocol, requiring only half an hour of analysis for each hour of videotape recording [90]; however, this "discount" approach does not provide enough time to even listen to the entire recording, and will only reveal a user's thought processes that are readily apparent.

One of the disadvantages of concurrent verbalization is that the process may interfere with task performance [33, 92]. By asking users to perform another task (think aloud) in addition to their primary task, data gathered on their primary task might be compromised. The fast pace and time constraints associated with entertainment technology exacerbate this problem. Asking subjects to verbalize what they are thinking also interferes with their natural utterances. To avoid this issue while still getting the benefits of a think aloud protocol, participants can be asked to perform the think aloud protocol retrospectively using video replay. This method is referred to as a retrospective think aloud protocol [33, 92] and is very valuable as the user can make more extensive comments than when constrained by the primary task. One drawback of retrospective testing is the time commitment needed to replay the task situation. With many specialized user groups (e.g., doctors, lawyers), the time factor would impede the use of this test; however, with some user groups (e.g., university students, computer game players) the benefits of using the technique outweigh the time commitment. Also, retrospection may lose some fidelity that would be present when discussing the task in real time.

2.3.2 Discount Techniques

There is significant overhead in terms of time and personnel required for an extensive empirical evaluation of software or an ethnographic study of users' habits and patterns of activity. Due to this overhead, a set of evaluation techniques called discount or low cost methods was introduced [70]. Used mainly for traditional usability testing, discount methods have become popular, but do not address some of the deeper issues that can be uncovered with a more formal investigation. Usability inspection [70], a type of discount technique, is the generic name for a set of methods anchored in having reviewers inspect or examine aspects of an interface related to usability. Two of the most popular inspection techniques are cognitive walkthrough and heuristic evaluation.

2.3.2.1 Cognitive Walkthrough

Many users prefer to learn about the functionality of a piece of software as the need arises, rather than through formal training. One feature of this approach is that the overhead invested in learning a new feature or task gives immediate benefit to the user. Cognitive walkthrough [68, 106, 142] is a usability inspection method with the goal of evaluating an interface for ease of learning through exploration [142]. The complex interactions between the cognitive processes of the user, the characteristics of the task, and the details of the interface create the processes through which a user learns a system [68]. Using a list of questions to focus their attention on the aspects of the interface that are important in facilitating the learning process, reviewers evaluate the interface in the context of a specific user task [68]. In a test, cognitive walkthrough detected almost 50% of the usability problems uncovered with a full-

scale evaluation [68], yet only took a fraction of the time. However, evaluators who were familiar with the theory of exploratory learning found more agreement and observed more error paths than evaluators unfamiliar with the theory [68].

Recently, Pinelle and Gutwin [104, 105] adapted the method of cognitive walkthrough for use with groupware systems. In groupware walkthrough, reviewers step through the tasks with the intention of evaluating how well the interface supports teamwork. The technique can be applied at any stage of the iterative design cycle, from low-fidelity prototypes to functioning applications [105]. Pinelle and Gutwin introduced the mechanics of collaboration [45, 105], a breakdown of the components of teamwork that support group members in working towards a shared outcome.

2.3.2.2 Heuristic Evaluation

Heuristic evaluation [70, 91, 93] is one of the most informal methods of usability inspection. It involves having usability specialists judge whether each interface element is consistent with established usability principles called heuristics [70]. Heuristic evaluation has been promoted as a cheap and quick method of identifying usability problems [93]. A set of evaluators should be used because a single individual will not be able to identify all of the usability problems in an interface. In fact, averaged over six projects, single evaluators only found 35% of the usability problems [91]. Through a review of the methodology in a number of studies, Nielsen recommends using 3-5 evaluators to identify most of the usability issues, and states that 5 evaluators will uncover 80% of the problems [91, 93].

Although Molich and Nielsen [86] identified a standard set of heuristics (discussed in detail in [92]), other heuristics can be used depending on the interface, application domain, and intended set of users. For example, Baker et al. [4, 5] developed a heuristic evaluation methodology for shared workspace groupware based on Gutwin and Greenberg's mechanics of collaboration [45]. Showing similar performance results to Nielsen's traditional heuristic evaluation, the groupware heuristics evaluate teamwork (the work of working together), in addition to taskwork [5].

A decade before Nielsen presented heuristic evaluation as a means of finding usability problems in productivity applications [91], Malone suggested a number of heuristics to make productivity software enjoyable to use, based on his evaluation of children playing different versions of an educational game [73]. These heuristics were organized into themes of challenge, fantasy, and curiosity. Game design has changed immensely since Malone's heuristics, and his choices were not made with the intention of evaluating games, but with the purpose of learning lessons from game design to apply to the design of productivity systems. Recently, there has been a renewed attempt to design heuristics specific to the domain of games [24, 34].

In 2002, Federoff [34] created a list of heuristics informed by a case study at a game development company. Federoff's heuristics were broken into three themes: interface (controls and display), mechanics (interacting with the game world), and gameplay (problems and challenges). She compared her heuristics to current game industry guidelines and Nielsen's heuristics, [93] and found that although Nielsen's heuristics encompassed many of the game interface issues, there were issues specific to

playability that were missing. Desurvire [24] introduced Heuristic Evaluation for Playability (HEP), a comprehensive set of heuristics for playability. HEP was based on productivity literature and playtesting heuristics that were specifically tailored to evaluate video, computer, and board games. An evaluation of the effectiveness of HEP showed the heuristics to be most salient for uncovering general issues in the early stages of development, using a prototype or mock-up.

More recently, Sweetsner and Wyeth [127] used heuristics to create a model for player enjoyment in games based on Csikszentmihalyi' s [21] concept of flow, which refers to optimal experience due in part to the appropriate balance between the skill of the participant and the challenge of the activity (see Section 3.5 for more detail on flow). Sweetsner and Wyneth's model, GameFlow [127], consists of eight elements: concentration, challenge, skills, control, clear goals, feedback, immersion, and social interaction. Each element includes a set of criteria for achieving enjoyment in games. Industry experts, using the strategy games Warcraft III and Lords of Everquest, evaluated the GameFlow model to expose weakness, ambiguities, or other problems with the model. The ratings provided by the evaluation matched fairly well with average ratings provided by professional game reviewers. Like other heuristic evaluation methods, the GameFlow model provides qualitative information on the enjoyment criterion (heuristics), as well as a rating scale for each element of the model.

2.3.3 Our Experiences with Current Methodologies to Evaluate Interactive Play

Since there have been no commonly used objective techniques for determining whether a certain technology creates an enjoyable experience, we have previously used many of the methods discussed in this chapter to evaluate interaction with play technologies.

We have used questionnaires extensively to gather user preference responses to different technological environments. For example, we used this technique to determine whether children (aged 11-13) preferred playing together on the same shared computer, side-by-side on separate computers, or on separate computers connected by a network [120]. We created and used child-friendly questionnaires, asking the children to rate the ease of the game on a scale from 1 to 5, where 1 corresponded to 'easy'; and 5 corresponded to 'hard'. We were able to determine that children found the game easier to play in the shared-display setting (mean = 2.3, S.D. = 0.8), compared to the side-by-side (mean = 2.8, S.D. = 0.8) or the separated (mean = 2.9, S.D. = 0.8) displays settings, (χ^2= 10.7, $p < 0.01$, Friedman two-way ANOVA). Although the children found it *easier* to play in the shared-display setting, we found that students did not always *prefer* playing in the shared-display setting. On interface evaluations conducted after each display configuration, children rated all three settings as being somewhat fun on a five-point scale, where one corresponded to 'fun', and five corresponded to 'not much fun' (shared: mean = 2.4, S.D. = 1.4; side-by-side: mean = 2.6, S.D. = 1.5; separated: mean = 2.6, S.D. = 1.4). On the post-experimental

questionnaire, when asked to choose which setting was the most fun to play, their preferences varied (shared: 30%, side-by-side: 25%, separated: 45%, $\chi^2 = 1.3$, *ns*).

In another research project, we designed an interface that semantically partitioned data over a number of handheld computers for children (pre-teen) to use while playing a game that helped them learn genetics concepts [22, 75]. Afterwards, the children filled out a post-session questionnaire. All seven participating students reported that they would prefer to play the game with a friend than by themselves. The children reported overwhelmingly that the face-to-face component was their favorite part of the experience. All seven children were extremely positive; six of the children ranked their enjoyment as either a four or a five on a five-point scale and the remaining child ranked their enjoyment a three [75]. Although the questionnaires provided numerical data concerning the children's enjoyment of the game, explanations were needed to elucidate their opinions. In many questionnaires using numerical scales, places are provided to explain the choices made [122]. However, it is sometimes difficult for participants to verbalize what aspect of a certain experimental condition they found less fun, challenging, or interesting.

Although there is a substantial time commitment, we have used observation analysis of video data to determine children's engagement when playing the same game in a paper condition, on a computer with one mouse, or on a computer with multiple mice input [120]. In the behavioural tradition, we recorded the play sessions, coded the events, and analyzed the results quantitatively. Our analysis included the amount of time in which the children played synchronously, the amount of time children engaged

in off-task behaviour, the amount of time children were inactive, and the children's physical pointing behaviour. The results of an ANOVA showed that the children exhibited significantly more off-task behaviour during the one-mouse computer condition (mean = 43.8 secs., S.D. = 70.0 secs.) than in the two-mice computer condition (mean = 13.1 secs., S.D. = 32.1 secs., $F_{1,32} = 9.835$, $p < 0.01$). In addition, a repeated-measures ANOVA showed a significant difference between the average inactivity across collaborative settings ($F_{2,26} = 123.51$, $p < 0.001$). A Tukey's HSD posthoc test showed that there was significantly more time in the one-mouse setting (mean = 374.6 secs., S.D. = 22.0 secs.) when both partners were inactive than in either the paper-based setting (mean = 195.4 secs., S.D. = 57.3 secs., $p < 0.05$), or the two-mice setting (mean = 173.4 secs., S.D. = 27.4 secs., $p < 0.05$).

We also used observational analysis techniques in the social tradition to examine the impact of display configuration on children's enjoyment playing a computer game while sharing a single display, sitting side-by-side with separate displays, or being separated by a network [119, 120]. In addition to event data (e.g., looking at the partner's screen), we coded verbal data including clarification statements, deictic references, and other conversational components. This enabled us to analyze the processes by which students interacted in these various collaborative settings, not simply the variance between the settings. Analysis of the conversations showed that the children sometimes had trouble reaching a mutual understanding of the workspace when using individual displays. The same conversational patterns were not present when sharing a display.

2.3.4 Where Current Methods Fall Short

Although we have previously used many of the methods discussed in this chapter, they all have limitations for understanding user experience with entertainment technologies. Our motivation is to evaluate traditional entertainment environments, but also to evaluate emerging play environments. This includes understanding how emerging technologies can enhance a player's experience with entertainment technologies, and how people respond to the inclusion of emerging technologies in their play environments.

Due to the market success of computer and video games (see section 1.1), there has been recent interest in using traditional methods to evaluate the playability of games, and to adapt traditional methods when they fall short. The evaluation of games requires a different set of tools than the evaluation of productivity systems because the ultimate goals of these domains are fundamentally different.

Pagulayan et al. [96], discuss nine characteristics in which games differ from productivity applications, and how these differences impact the choice of evaluation methodology. For example, the design intentions behind most productivity applications are to make tasks easier and quicker, to reduce errors made, and to increase the quality of the result. Evaluation of productivity systems focuses on producing a better result, and the process of using a well-designed application enhances the result. On the other hand, games are intended to be fun to play. The goal is to stimulate thinking and feeling, and the result of a game serves to enhance the pleasure of the *process* of playing. This fundamental difference between an emphasis

on result or process is just one of the many ways in which games and productivity applications differ [96]. A consequence of these differences is that the traditional techniques for evaluating productivity applications that are outlined in this chapter may fall short when used to evaluate entertainment technologies.

Subjective techniques such as questionnaires are good approaches to understanding the *attitudes* of the users, but subjects are bad at self-reporting their *behaviours* in game situations [43]. As previously discussed, since subjective reports are cognitively mediated, they may not correspond to the actual experience [79, 149]. In addition, participants' reaction to new play environments might be skewed by the novelty of the entertainment technologies. Although subjective techniques are a good approach to understand *user preferences*, these techniques do not uncover much information on *user behaviours*.

Task performance is a widely used metric in HCI, when improved productivity and performance are the goals of the technology. Performance metrics are not particularly useful for evaluating play technologies since the success of an entertainment technology is not related to the *performance* of the participant, but to the *experience* of the participant. A player can have a very enjoyable play experience while losing a game, and can also be bored with an overwhelming win. In play, the process is more important to success than the result [96]. As such, task performance is not a very useful metric for evaluating user experience with play technologies.

Observational techniques including verbal transcriptions, gesture analysis, and facial expression analysis can provide insight into a play experience. These observations

can be analyzed as they occur within the play context, grounding the data within the experience. However, the tremendous time commitment, and the need for specialized training renders observational analysis impractical for many researchers, while developers of play technologies are completely prohibited by time and budget. There are a few consulting firms (e.g., [151]), that specialize in observational analysis of entertainment technologies, but game companies may pass up these services since shipping dates of games takes priority over the evaluation of playability. There has been recent interest in observational usability testing methods in order to access information about user behaviour [41], but little testing has been performed to determine the efficacy of usability testing for games.

Standard discount usability methods, such as heuristic evaluation and cognitive walkthrough are useful for finding where there are breakdowns between a user's cognitive model of how a system functions, and the actualization of the system. Although useful for uncovering usability issues within play environments, there has been minimal comparable research on using heuristics to evaluate the playability of an entertainment technology [24], or to evaluate the impact of the introduction of an emerging technology on user experience. Most importantly, these discount methods do not involve actual users, but are administered by specialists in the domain of usability. When research involves incorporating novel technologies into a play experience, there are no "experts" who can use their expertise to determine how a regular user will feel. At this point, researchers can only guess how the technologies will impact the users.

Traditional evaluation methods have been adopted, with some success, for quantitative-subjective, qualitative-subjective, and qualitative-objective assessment of entertainment technologies. When evaluating productivity systems, metrics of task performance are used for quantitative-objective analysis (see Figure 1), but as previously mentioned, task performance is not very relevant when evaluating entertainment technologies. As such, there is a knowledge gap for quantitative-objective evaluation of entertainment and play (see Figure 2) and a new evaluation methodology is needed.

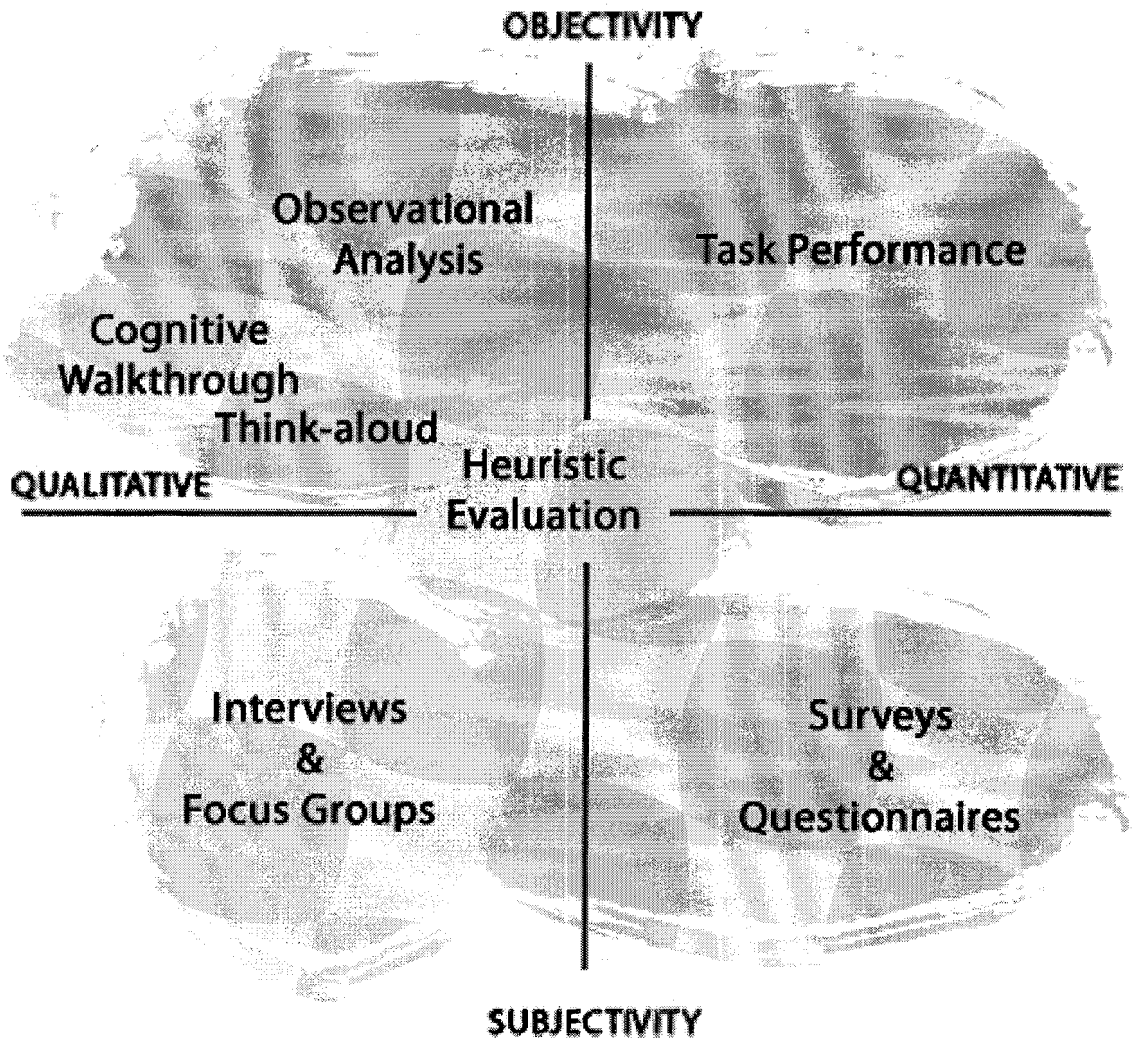


Figure 1: Current methods for evaluating productivity applications. Evaluators have a lot of choice and can pick the evaluation method that best suits their needs. Note that heuristic evaluation can be seen as a quantitative methodology since experts can provide ratings for how well software adheres to the heuristics. Observational analysis is a tool that can be used to generate quantitative or qualitative results.

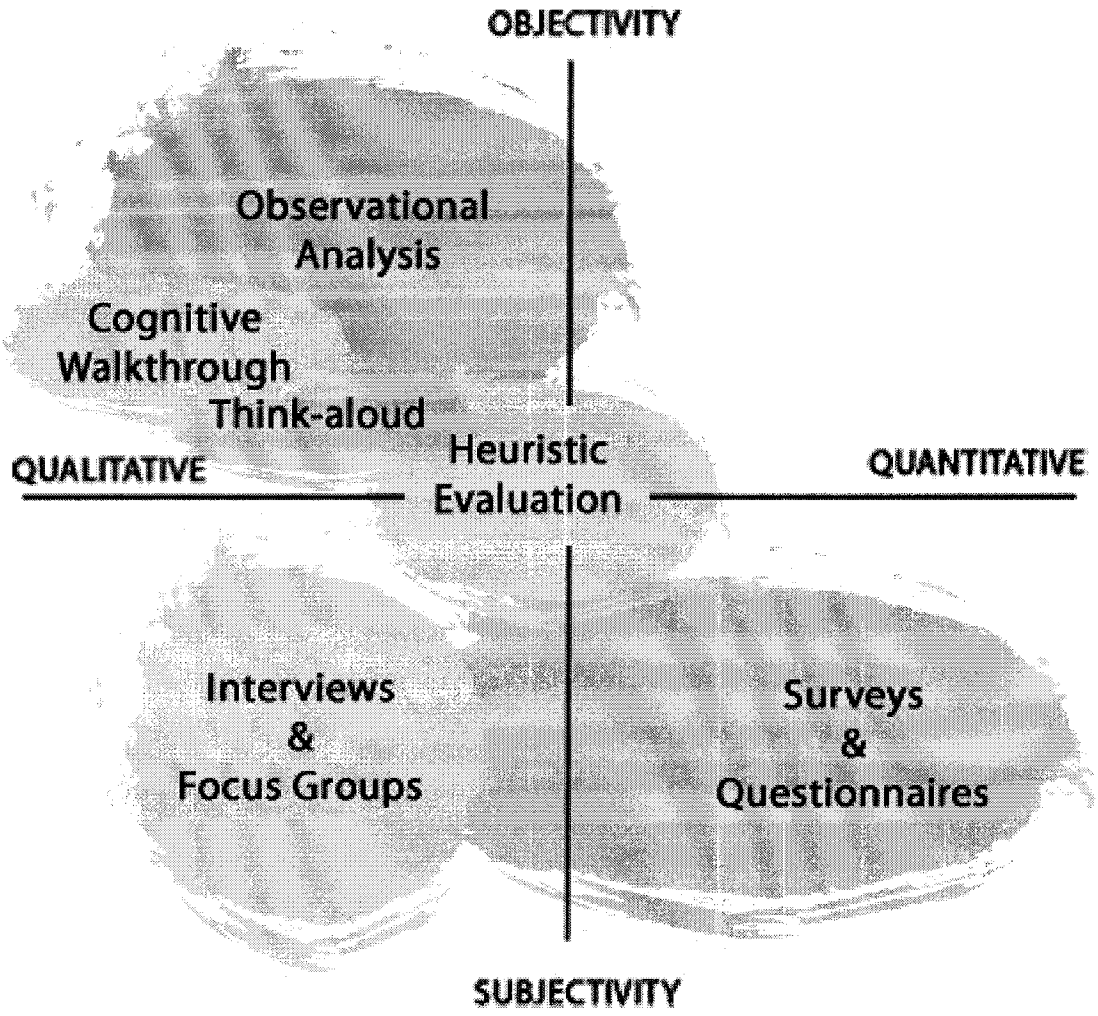


Figure 2: **Current methods for evaluating entertainment technologies. Evaluators have a lot of choice, but there is a knowledge gap in the quantitative-objective quadrant since task performance metrics aren't relevant. Heuristic evaluation can be seen as a quantitative methodology since experts can provide ratings for how well software adheres to the heuristics. Observational analysis is a tool that can be used to generate quantitative or qualitative results, but is not used quantitatively to evaluate entertainment technologies due to the time commitment and expertise needed.**

Chapter 3 RELEVANT PSYCHOLOGICAL CONCEPTS

Chapter 2 shows how current evaluation methods fall short for evaluating play technologies. In the next three chapters, we will present literature that supports the idea that physiological signals from the body can be used to generate a new, objective and quantitative evaluation methodology, fit for evaluating interaction with play technologies.

Before discussing the physiological measures and how to apply them to an evaluation methodology, it is important to identify and describe some of the most important physiological and psychological concepts related to this area of research. The nervous system is described first, followed by the psychological concepts of arousal, valence, and habituation, which are central to psychophysics research. An introduction to the study of affect and emotion is also provided.

3.1 Human Nervous System

The nervous system (see Figure 3) is divided into two components: the central nervous system (CNS) which consists of the brain and spinal cord, and the peripheral nervous system (PNS), which is composed of all of the ganglia and nerves that lie outside of the CNS [60]. The PNS carries information between the body and the CNS [80]. This information is either gathered from sensory receptors and sent to the CNS (afferent

division), or can be output from the CNS to the body (efferent division). The PNS is divided into the somatic nervous system and autonomic nervous system (ANS) [80]. The somatic nervous system is a voluntary system that controls the skeletal muscles for body movement and provides information to the CNS on muscle and limb position [60]. The ANS (sometimes called the visceral nervous system) controls actions in the body that we do not have conscious control over, including smooth muscle control, cardiac muscles, and glandular activity. The ANS regulates body temperature, and coordinates cardiovascular, digestive, respiratory, excretory, and reproductive functions [80].

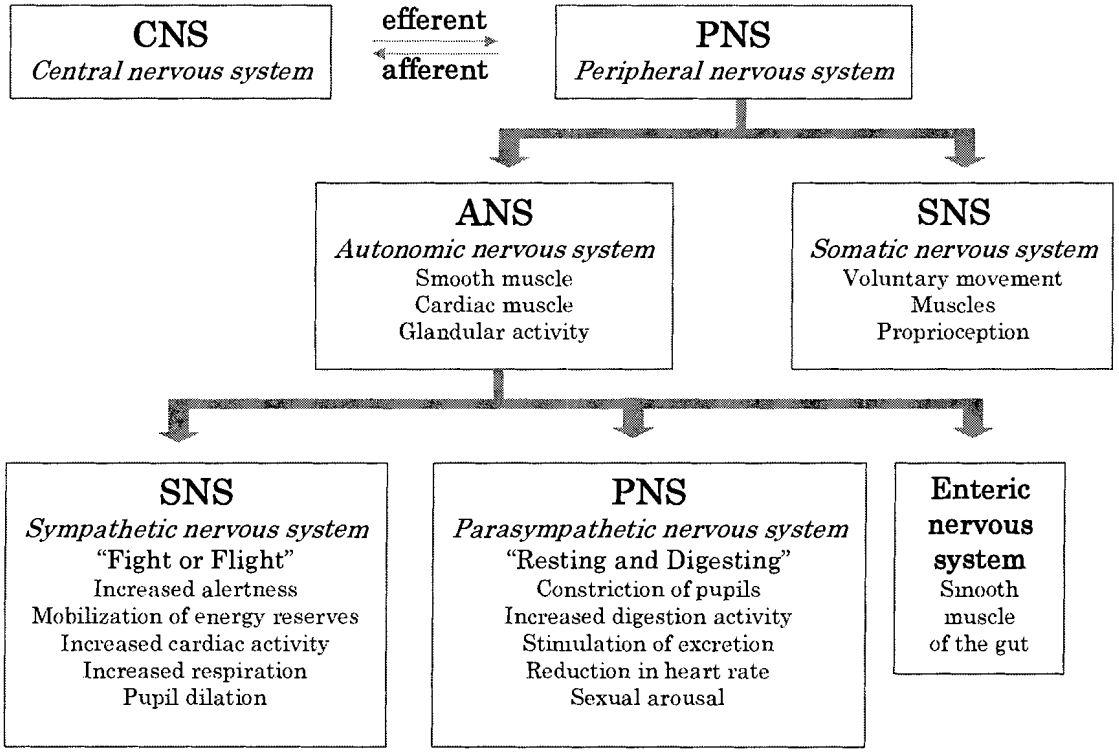


Figure 3: The Human Nervous System.

The ANS is further subdivided into three anatomically separate branches, the sympathetic nervous system (SNS), the parasympathetic nervous system (PNS), and the enteric branch [28, 60], which is responsible for controlling the function of the smooth muscle of the gut [60]. We will focus on PNS and SNS activity. The SNS has dominant function in emergency situations and is used in "fight or flight" situations, such as athletic competition, combat, severe temperature changes, and blood loss [28]. SNS activation causes us to experience increased alertness, a feeling of energy, increased activity in the cardiovascular and respiratory systems, pupil dilation, and the mobilization of energy reserves [28, 80]. The PNS is the relaxed activity controller and is responsible for activities such as resting and digesting. Also called the anabolic system, effects produced by the PNS include: constriction of the pupils; increased muscle and glandular activity related to digestion; stimulation and correlation for excretion; reduction in heart rate; and sexual arousal [80]. Under normal circumstances, there is a balance between the PNS and SNS systems.

3.2 Arousal

The concept of arousal stems from Cannon's theory of the unified body preparing for fight or flight [123]. It is described as "a state of heightened physiological activity" [26]. Although used extensively by psychologists as a means of describing psychological activity, the concept of arousal is still grounded in physiological changes in the body.

3.3 Valence

The psychological definition of valence is "the degree of attraction or aversion that an individual feels toward a specific object or event" [26]. Valence describes where an emotional reaction sits on an axis from the positive to the negative.

3.4 Habituation

Where arousal suggests a heightened response to a stimuli, habituation refers to the reduction of response based on exposure to previous and repeated presentation of the same stimulus [123]. Habituation can be differentiated into short and long-term habituation, where short-term occurs within a short period of time, such as within a single testing session. Long-term habituation can occur over days or even weeks. Habituation is important to consider, because as participants of a study are exposed to a stimulus, it is possible that their responses (both overt and autonomic) to the same stimuli will adapt over the course of a session.

3.5 Flow

Csikszentmihalyi [21] was interested in what makes experiences enjoyable, and conducted extensive research over a decade, collecting survey and interview data from several thousand participants all over the world. He discovered that optimal experience, which he labeled as 'flow', is the same for very different tasks, and that flow transcends culture, social class, age, and gender.

Flow refers to an experience state that causes deep enjoyment, due in part to the right balance between the skill of the participant and the challenge of the activity.

Csikszentmihalyi developed a set of eight elements that contribute to a state of flow, including: (1) a task that can be completed; (2) the ability to concentrate on the task; (3) that concentration is possible because the task has clear goals; (4) that concentration is possible because the task provides immediate feedback; (5) the ability to exercise a sense of control over actions; (6) a deep but effortless involvement that removes awareness of the frustrations of everyday life; (7) concern for self disappears, but sense of self emerges stronger afterwards; and (8) the sense of the duration of time is altered [21].

3.6 Emotion

Emotions have historically been examined from two perspectives:

1. Emotions are cognitive, stressing their mental component
2. Emotions are physical, stressing their physical component [101].

The former theory can be traced to Cannon, who believed that emotion is experienced by the brain and that emotion is possible without sensations from our bodies (Cannon, 1927 as cited in [101]). He also provided evidence that autonomic events are too slow, too insensitive, and not distinct enough to contribute to emotions [14]. The original proponent of the physical theory of emotion was William James. James emphasized that emotion was experienced as bodily changes, such as sweating hands and a fast beating heart (James, 1890 as cited in [101]). Not only did he maintain that discrete emotional experiences could be identified by unique patterns of bodily changes, he believed that the perception of these physiological changes *is* the emotional experience [14]. Recent research has shown that the answer lies

somewhere between these two extremes, as both brain and body can shape human experience. Thoughts and relived memories can elicit an emotional experience, as can changes in our body chemistry. Although our present research is not concerned with the underlying theory of emotional experience, it is important to understand the historical perspective as many of the recent efforts in using ANS activity as an indicator of experience were inspired by James's theory.

Chapter 4 PHYSIOLOGICAL MEASURES AND EMOTION

In 1964, John Stern defined psychophysiology as any research in which the dependent variable (the subject's response) is a physiological measure and the independent variable (the factor manipulated by the experimenter) a behavioural one [123]. Recent work has shown this view to be limiting since it is equally likely that we could manipulate physiological variables and view the effect on the psychological variables. Thus, the modern definition states that psychophysiology is effectively the study of the interaction between mind and body [123].

In medical fields, biofeedback is a technique whereby patients receive feedback about their physiological state to learn to control some aspect of their health [95]. For example, physiological indicators of tension and stress may be presented to patients who suffer from stress disorders, panic attacks, and hyperventilation. Biofeedback can help patients manage their stress by prompting them to engage in breathing exercises or other stress reduction techniques during times when their stress levels are too high. Patients can then see their stress reducing as a direct result of the stress-reduction techniques employed.

This section presents information relevant to the physiological measures used in biofeedback or psychophysical research. Organized by anatomical system, each

subsection presents: the measure; its psychological counterpart; other factors it is affected by; devices used to measure it; and references of its use.

4.1 Skin: Electrodermal Activity

Electrodermal activity refers to the electrical properties of the skin. Also called the galvanic skin response (GSR) or the psychogalvanic reflex, it is easily measured as either skin resistivity or skin conductance. This choice has implications for the interpretation of results [13]. Electrodermal activity is one of the most commonly used physiological responses in psychophysiological research and in computing systems that integrate body responses.

There are two components to the electrodermal response: the tonic baseline and the short term phasic responses superimposed on the baseline [123]. The phasic response is called the electrodermal response (EDR), skin conductance startle response (usually as a response to extreme stimuli), skin conductance orienting response (general term), skin resistance response (SRR), or the skin conductance response (SCR) [10]. It is thought that the electrodermal response evolved for locomotion, manipulation and defense [123]. There are specific sweat glands, called the eccrine sweat glands, which are used for measuring GSR. Located in the palms of the hands and soles of the feet, these sweat glands respond to psychic stimulation instead of simply to temperature changes in the body. For example, many people have cold clammy hands when they are nervous. In fact, subjects do not have to even be sweating on the palms of the hands or soles of the feet to generate differences in skin conductance because the eccrine sweat glands act as variable resistors on the surface. As sweat rises in a

particular gland, the resistance of that gland decreases even though the sweat may not overflow onto the surface of the skin [123].

4.1.1 Psychological Counterpart

Galvanic skin response is a linear correlate to arousal [63] and reflects emotional responses as well as cognitive activity [10]. GSR has been used extensively as an indicator of stress and mental workload in both non-technical domains (see [10] for a comprehensive review), and technical domains. It is considered the most sensitive response used in the detection of deception (lie detectors) [10] and has also been used to differentiate between anger and fear [12].

Although electrodermal activity is widely recognized in psychophysiology, there are other factors that affect the galvanic skin response including age, sex, race, temperature, humidity, stage of menstrual cycle, time of day, season, sweating through exercise, and deep breathing [10, 123, 136]. There are also individual differences stemming from personality traits such as whether an individual is introverted or extroverted [10]. Due to these differences, it is difficult to compare GSR across groups of individuals or in the same individual across different test sessions. In a single session, skin conductance does not have to be corrected for base level, whereas skin resistivity does [123].

4.1.2 Devices and Use

Devices used to measure GSR range from simple circuits attached to aluminum foil finger cuffs to high-end systems used to detect deception. Wearable devices, devices

that are embedded into clothing or accessories, have recently been designed to decrease interference from bulky equipment. The MIT Media Lab has designed a glove called the galvactivator [42], GSR rings and bracelets [1], GSR shoes [1], and a standard skin sensor [85]. A brief visit to the Lego Mindstorms community bulletin boards [66] revealed a few instances of using Lego components to build simple lie detectors using GSR.

Figure 4: Galvanic skin response (GSR) was collected using surface electrodes that snap onto Velcro straps worn around the index and ring fingers.

We measured GSR using surface electrodes sewn in Velcro straps that were placed around two fingers on the same hand (see Figure 4). Previous testing of numerous electrode placements was conducted to ensure that there was no interference from movements made when manipulating the game controller (see Appendix 3). We found that finger clips were as responsive to GSR as pre-gelled electrodes on the feet, while electrodes on the palms suffered from movement artifacts.

GSR feedback has been used in the medical community for relaxing and desensitization training, and in the treatment of excessive sweating (hyperhydorses) and related dermatological conditions. As an input to interactive systems, GSR has been used in the Relax-to-Win racing game [7], in the analysis of driver stress [48], in an instant message application called Conductive Chat [25], for an automated music selection DJ [46], and in the Conductor's Jacket [78] (see Chapter 5.6 for a more detailed description of these systems).

4.2 Cardiovascular System

The cardiovascular system includes the organs that regulate blood flow through the body. Measures of cardiovascular activity include heart rate (HR), interbeat interval (IBI), heart rate variability (HRV), blood pressure (BP), and blood volume pulse (BVP). Heart rate indicates the number of contractions of the heart each minute, while HRV refers to the oscillation of the interval between consecutive heartbeats. Blood pressure is a measure of the pressure generated to push blood through the arteries, veins, and capillaries, while BVP refers to the amount and timing of blood flowing through the periphery of an individual.

4.2.1 Blood Pressure

Blood pressure indicates how much pressure is needed to push blood through the system of arteries, veins, and capillaries. Although blood pressure is known to be affected by age, diet, posture, and weight, it is also affected by the setting (clinical vs. normal) and by highly stressful situations [123]. Generally, BP is collected using an inflated arm cuff (sphygmomanometer) that is inflated, and subsequently deflated

while readings are taken. As a result of cuff inflation and deflation, blood pressure responses to stimuli cannot generally be collected in real-time. There were some sophisticated and expensive pieces of equipment that were developed to collect BP continuously, but these systems were removed from the market due to their lack of commercial success. Automated machines have been developed for use with polygraph machines, but cannot accurately take more than one reading per minute [123].

4.2.2 Blood Volume and Pulse Volume

Blood volume reflects slow changes in the tonic level of an appendage while pulse volume is a phasic measure of the pulsatile change in blood flow related to both the pumping of the heart and to the dilation and constriction of blood vessels in the periphery [123]. Thus, pulse volume (BVP) measures the amplitude of individual pulses. BVP increases in response to pain, hunger, fear and rage and decreases in response to relaxation [123]. BVP is difficult to collect outside of a clinical environment because it is affected by room temperature and is very sensitive to placement and motion. Due to these same factors, comparison between subjects is not possible.

4.2.2.1 Devices and Use

BVP is collected using a plethysmograph. Photoelectric plethysmography uses a photocell placed over an area of tissue (e.g., finger). A light source is passed through the tissue (or bounced off the tissue), and the amount of light passed through (or bounced back) is measured by a photoelectric transducer [123]. Impedance

plethysmography employs two electrodes through which a high-frequency alternating current is passed. Changes in blood volume affect the electrical impedance giving a reading of BVP [123]. Strain gauge plethysmography uses a strain gauge placed around the finger or toe. Changes in resistance or voltage of the strain gauge can be considered an indirect measurement of blood volume [123]. Venous occlusion plethysmography uses two inflated cuffs on the same appendage. As with BP measurements, since cuffs are used, real time measurements are not possible [123].

BVP is generally collected using the finger or toe. Since blood pulses through the earlobe, one might think that the earlobe is a convenient location to measure BVP. However, for BVP measurements, the earlobe is not as responsive as the finger to typical laboratory tasks [123]. When responding to stimuli, the body prepares for fight or flight by increasing or decreasing blood flow to the peripheral organs. The earlobe is not one of the places that selective increases in blood flow occur. We did not collect BVP in any of our experiments because the sensing technology used on the finger is extremely sensitive to movement artifacts. As our subjects were operating a game controller, it wasn't possible to constrain their movements.

4.2.3 Heart Rate

Heart rate (HR) indicates the number of contractions of the heart each minute, and can be gathered from a variety of sources. HR has been used to differentiate between positive and negative emotions, with further differentiation made possible with finger temperature [97]. Distinction has been made in numerous studies between anger and fear using HR [97] (for a comprehensive review, see [12]).

In addition to the psychological differences that HR elicits, it is also affected by age, posture, level of physical conditioning, breathing frequency, and circadian cycle[3].

4.2.4 Heart Rate Variability

Heart rate variability (HRV) refers to the oscillation of the interval between consecutive heartbeats (IBI). The heart rate of a normal subject at rest is irregular. This irregularity is called sinus or respiratory arrhythmia [59]. Fluctuations around the mean heart rate are respiratory-related, baroflex-related[4], and thermoregulation-related. We are most concerned with the baroflex-related fluctuation. Blood pressure changes are detected by baroreceptors in the aorta. An increase in blood pressure causes a sympathetic inhibitory response, and in turn, the effects of this response cause a decrease in blood pressure, creating a negative feedback loop [133]. The passage of the neural signal from the baroreceptors through the brainstem is associated with a time delay of about 1 sec [87]. This time delay creates a phase shift and causes the system to oscillate. The oscillation frequency is about 0.1 Hz [87]. If IBI is fairly constant, then HRV will be low, whereas if IBI is changing (regardless of absolute value), then HRV will be higher.

In 1963, Kalsbeek and Ettema [58] found a gradual suppression of heart rate irregularity due to increasing task difficulty. Later, Kalsbeek and Sykes [59] tested a motivated group versus a non-motivated group (using money as a motivator), and found that the motivated group maintained a constant level of suppression while the

---

[3] Relating to or exhibiting approximately 24-hour periodicity [26].
[4] Baro is relating to pressure [26].

non-motivated group started at a lower level of suppression and continued to decline. Since then, many researchers have attempted to use HRV as an indicator of mental effort.

HRV has been used extensively in the human factors literature as an indication of mental effort and stress in adults. In high stress environments such as ambulance dispatch [141] and air traffic control [113], HRV is a very useful measure. When subjects are under stress, HRV is suppressed and when they are relaxed, HRV emerges. Similarly, HRV decreases with increases in mental effort [113] and cognitive workload [144], but as the mental effort needed for a task increases beyond the capacity of working memory, HRV will increase [110, 113]. Many researchers have found significant differences in HRV as a function of mental workload (see Chapter 5), while others have not [83, 87]. HRV has also been used to differentiate between epistemic behaviour (concerning the acquisition of information and knowledge), and ludic behaviour (playful activities which utilize past experience) in children [53].

One method of determining HRV is through a short-term power spectral density analysis of interbeat interval, which is described in the next section.

4.2.4.1 Spectral Analysis of Sinusarrhythmia

Power spectral density analysis describes how power is distributed as a function of frequency. Using the interbeat interval (R-R interval on an EKG, see Figure 5), power spectral density analysis provides a measure of how the heart fluctuates as a result of changes in the autonomic nervous system [113]. The high frequency component (0.15-

0.4 Hz) is associated with parasympathetic nervous system activity (resting and digesting), while the low frequency component (0.04-0.15 Hz) is associated with sympathetic nervous system activity (fight or flight) [56, 83]. A ratio of the low frequency to high frequency energy in the spectral domain is representative of the relative influences of the sympathetic to parasympathetic influences on the heart.

Recently, researchers have used spectral analysis of sinus arrhythmia (heart rate variability) to provide an objective measure of mental effort[5]. Mulder proposed that controlled processing is required to: locate and maintain information in short term memory; retrieve information and programs from long term memory; and make decisions, and that the total amount of controlled processing is a function of the amount of effort a subject invests in a task [87]. Associated with controlled processing are spontaneous oscillations in blood pressure (and the cardiac interval signal) around 0.1 Hz [87]. A loss of the 0.1 Hz frequency component would decrease the variance and thus suppress heart rate variability. Measuring HRV using the 0.1 Hz frequency component has the important advantage of being able to discriminate between the effort-related blood pressure component, and the effects due to respiration, motor activity, and thermoregulation, since these other factors influence other parts of the power spectrum [137].

In order to perform spectral analysis, researchers used to convert the interval signal to an equidistant time series using interpolation, or filtering [87]. Recent digital

---

[5] Note that mental effort does not seem to vary as a function of mean heart rate, but of the variability of heart rate [83].

technology produces a measure of the interbeat interval at 4Hz, which can be used directly. This time series data is then smoothed and Fourier-transformed. The frequency range sensitive to changes in mental effort is between 0.06 and 0.14 Hz [137], while the area between 0.22 and 0.4 Hz reflects activity related to respiration [56, 87]. Integrating the power in the band related to mental effort provides a measure of HRV. Vicente recommends normalizing the measure by dividing by the average of all resting baselines and subtracting from one [137]. Then, a value between zero and one is produced where zero indicates no mental effort and one indicates maximum mental effort.

4.2.5 Electrocardiography

EKG (Electrocardiography) measures electrical activity of the heart. During each cardiac cycle, a wave of depolarization radiates through the heart [80]. This electrical activity can be measured on the body using surface electrodes. An example of an EKG signal is shown in Figure 5.

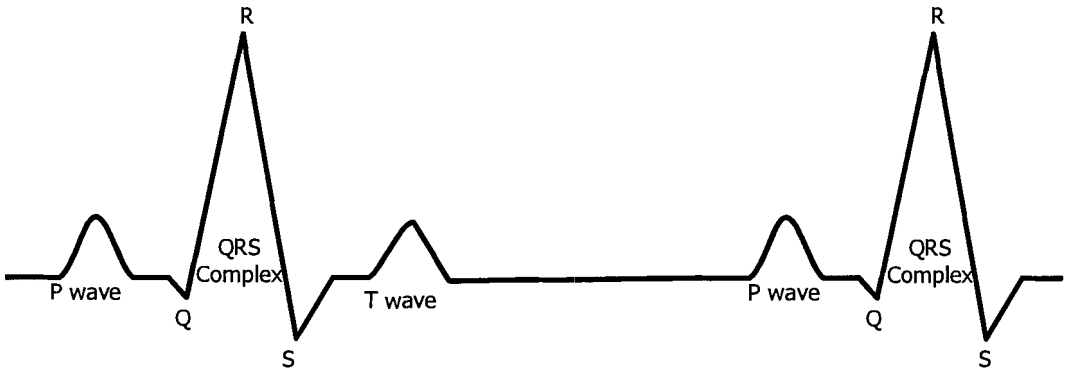


Figure 5: EKG signal. The P wave appears as the atria depolarize, the QRS complex accompanies the depolarization of the ventricles, and the T wave denotes ventricular repolarization. The R to R interval is the interbeat interval used to determine heart rate variability.

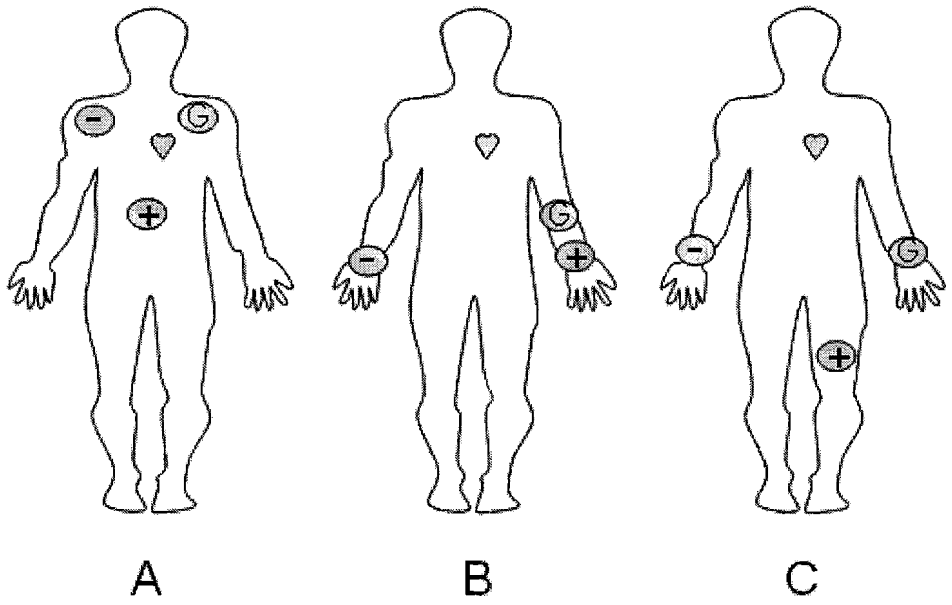


Figure 6: **Three common electrode placements for EKG. A) Chest placement. B) Forearm placement. C) Forearm and leg placement. (Adapted from Thought Technologies [130].)**

Heart rate, interbeat interval (IBI), HRV, and respiratory sinus arrhythmia (RSA) can all be gathered from EKG. Although there is a standard medical configuration for placement of electrodes, any two electrodes placed fairly far apart will produce an EKG signal [123]. The main placement method is on the chest with the negative electrode on the right shoulder, the positive electrode on the abdomen, and the ground on the left shoulder (see Figure 6: A), although the forearm provides a good measurement location for less intrusive measurement (Figure 6: B and C). EKG provides a good signal with which to derive the aforementioned physiological cardiac measurements.

Figure 7: EKG was measured using three surface electrodes, with two placed on the chest, and one placed on the abdomen.

We placed three pre-gelled surface electrodes (see Figure 7) in the standard configuration of two electrodes on the chest and one electrode on the abdomen (see Figure 6: A). Body hair can interfere with an EKG signal, and shaving the regions for electrode placement is a common clinical practice. As an alternative, we screened our participants to have little to no body hair on the chest or abdomen.

4.3 Respiratory System

Respiration can be measured as the rate or volume at which an individual exchanges air in their lungs. Respiration can be characterized by the following metrics: tidal volume (V_T), which is the volume that is displaced in a single breath; duration of inspiration; duration of expiration; and total cycle duration [143]. Minute volume (V_{MIN}) is calculated as the tidal volume divided by the respiration rate, and indicates

the volume that is displaced during one minute [143]. The commonly used measures in psychophysiological research are simply the rate of respiration and depth (amplitude) of breath [123].

Respiratory measures are most accurately measured by gas exchange in the lungs, but the technology inhibits talking and moving [123]. Instead, chest cavity expansion can be used to capture breathing activity using either a Hall effect sensor, strain gauge, or a stretch sensor [123]. In our experiments, we used a stretch sensor sewn into a Velcro strap, positioned around the thorax (see Figure 8).

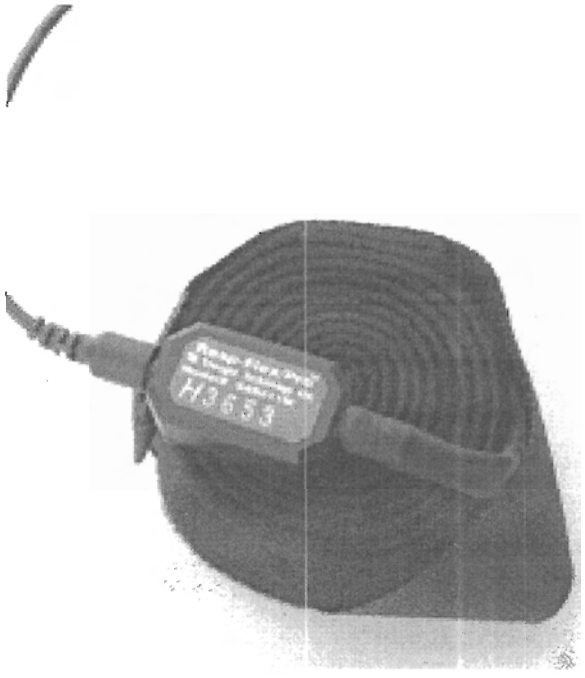


Figure 8: A stretch sensor was positioned around the thorax to measure respiration.

Emotional arousal increases respiration rate while rest and relaxation decrease respiration rate [123]. Although respiration rate generally decreases with relaxation, startle events and tense situations may result in momentary respiration cessation. Negative emotions generally cause irregularity in the respiration pattern [123]. In addition, states of pain, apprehension, anxiety and fear, threat and anger have been associated with hyperventilation [143]. Mental effort, stressful mental task performance, and high cognitive activity have been associated with an increase in respiration rate and V_{MIN}, and with a decrease in V_T, depth of respiration, and in the variability of respiration [143, 144]. Besides its psychological counterparts, respiration is affected by physical activity. Also, a deep breath can affect cardiovascular measures because respiration is closely linked to cardiac functioning.

Examples of the use of respiration measures are in the Conductor's Jacket [78], to quantify driver stress [48], and to measure stress in air traffic control simulations [134, 135].

4.4 Muscles: Electromyography

Electromyography (EMG) is the measure of muscle activity either through needles or surface electrodes. EMG measures muscle activity by detecting surface voltages that occur when a muscle is contracted [123]. Two electrodes are placed along the muscle of interest and a third ground is placed off the axis.

In isometric conditions (no movement) EMG is closely correlated with muscle tension [123]; however, this is not true of isotonic movements (when the muscle is moving).

When used on the jaw, EMG provides a very good indicator of tension in an individual due to jaw clenching [12]. On the face, EMG has been used to distinguish between positive and negative emotions [39]. EMG activity over the brow region (corrugator supercilii, the frown muscle) is lower and EMG activity over the cheek (zygomaticus major, the smile muscle) and preiocular (orbicularis oculi) muscle regions are higher when emotions are mildly positive, as opposed to mildly negative [12]. These effects are stronger when averaged over a group rather than for individual analysis, and have been able to distinguish between positive, neutral and negative valence at a rate greater than chance when viewing pictures or video as stimuli [99]. Tonic activity from EMG on the forehead (musculus frontalis, the eyebrow-raising muscle) has been used as a measure of mental effort [23, 39]. In addition to emotional stress and emotional valence, EMG has been used to distinguish facial expressions and gestural expressions [123].

EMG feedback is generally used for relaxation training, headache, chronic pain, muscle spasm, partial paralysis, speech disorder, or other muscular dysfunction due to injury, stroke, or congenital disorders.

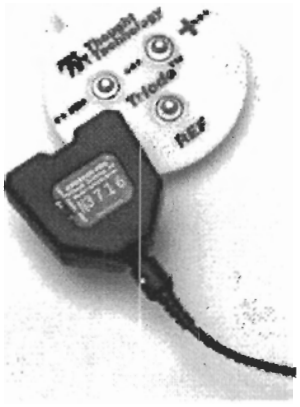


Figure 9: A preconfigured triangular arrangement of electrodes was used to collect EMG activity on the cheek and jaw.

In our experiments, we used surface electrodes to detect EMG on the jaw, (indicative of tension), and on the forehead (indicative of frowning), and cheek (indicative of smiling). On the jaw and cheek, we used three electrodes preconfigured in a triangular arrangement (see Figure 9). Due to the small size of the corrugator supercilli muscle, we used the extender cables (like the EKG electrodes seen in Figure 7) to collect EMG on the forehead. The disadvantage of using surface electrodes is that the signals can be muddied by other jaw activity, such as smiling, laughing, and talking. Needles are an alternative to surface electrodes that minimize interference, but were not appropriate for our experimental setting. Body hair can interfere with an EMG signal, and shaving the regions for electrode placement is a common clinical practice. As an alternative, we screened our participants to have clean-shaven faces in any of the regions where electrodes were to be placed.

4.5 Indexing Psychological Events From Physiological Data

> "Ever since psychologists started the study of bodily changes during emotion, there has been the hope that some patterns would turn up that would differentiate one emotion from another." (Woodworth and Schlosberg, 1954, as cited by [67]).

William James first speculated that patterns of physiological response could be used to recognize emotion [13], and although this viewpoint is too simplistic, recent evidence suggests that ANS activity can differentiate among some emotions [67]. There are myriad issues associated with this process, creating a difficult, poorly understood space that psychophysiologists must operate in. These issues and limitations will be discussed along with some potential solutions.

4.5.1 Classification of Emotion

There have been many methods proposed for classifying basic emotions. Researchers who adopt the idea of discrete, specific emotions hold that there are eight or nine basic, inborn emotions [27]. Emotions included in the set of basic emotions have varied. Ekman initially proposed seven distinct emotions (anger, disgust, fear, happiness, sadness, surprise, and contempt), but recently amended the list, adding: amusement, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure, and shame [29]. Although comprehensive, this list may not represent a typical emotional response to entertainment technology. One would expect excitement, pride, and satisfaction to play a role, but shame and guilt might be excluded. Lazarro [65] has qualitatively identified relevant emotions during game play including fear, surprise, disgust, naches (a Yiddish term for pleasure or

pride at the accomplishment of a child or mentee), fiero (an Italian term for personal triumph over adversity), schadenfreude (a German term for gloat over the misfortune of a rival), and wonder. Lazzaro has not made an attempt to measure these relevant emotions, and perhaps human experience states such as engagement, frustration, boredom and challenge are more salient descriptions than human emotions in our domain of study.

Another method of classifying emotion is by positioning an emotion along multiple axes in space, where the axes represent metrics of similarity. The arousal-valence space used by Lang [63] places stimuli in a 2-D space defined by arousal and valence (pleasure). Using pictures as stimuli, Lang and colleagues mapped individual pictures to arousal and valence levels.

Russell also used an arousal-valence space to create the Affect Grid. Based on their circumplex model of emotion, the affect grid is a tool to quickly assess affect along the dimensions of pleasure and arousal [114]. Subjects place checkmarks in the squares of the grid, as a response to different stimuli. Instead of only having two axes, the circumplex model has four axes including Stress-Relaxation and Depression-Excitement in addition to Arousal-Sleepiness and Pleasure-Misery (see Figure 10). Although the circumplex model uses four axes, emotions are still defined in two dimensions, arousal and valence.

One problem with the arousal-valence method of classifying mood is that arousal and valence may not be entirely independent and can impact each other. For example, Lang et al. [64] had difficulty finding images that represent the extreme regions of the

unpleasant/calm quadrant. It seems that if an image is truly unpleasant, it cannot also be calm, suggesting some interplay between these two axes.

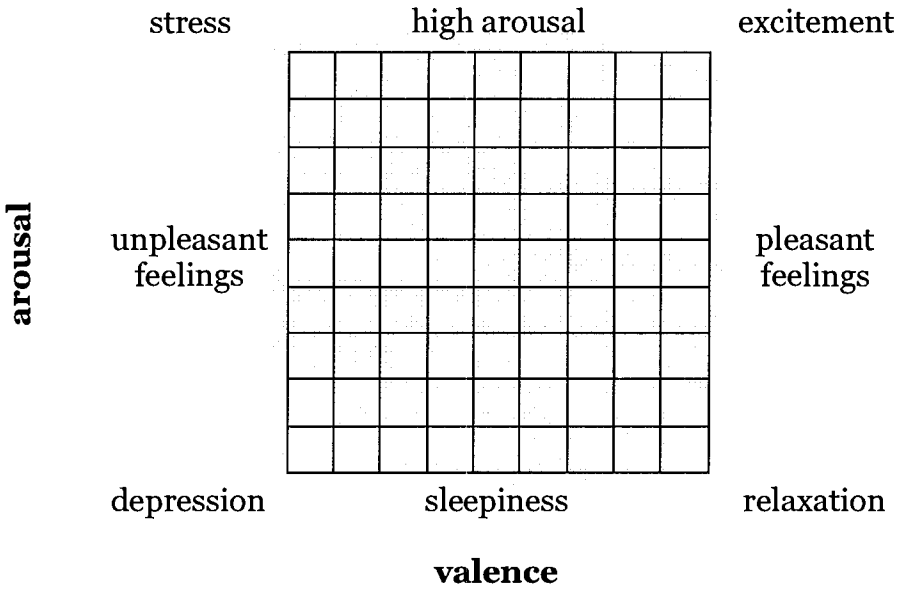


Figure 10: The Affect Grid: Based on the circumplex model of emotion, the affect grid allows for a quick assessment of mood as a response to stimuli [114]. Adapted from Russell et al. (1989) [114].

4.5.2 Issues and Limitations with Sensing Physiological Responses

Some of the criticism of James's theory can be attributed to issues and limitations with sensing patterns of physiological change. Ekman, Levenson and Friesen attribute many of the inconsistencies in this line of research to methodological problems [31]. These methodological problems include: the inability to isolate a single emotion; the failure to address the intensity of the emotion; the mistiming of ANS recording; and the need for simultaneous examination of a number of metrics [12, 14]. Beyond

methodological issues, there are issues inherent to examining emotions, and physiological correlates of emotional state.

4.5.2.1 Emotions

Emotions are very short-lived and typically last only for a few seconds [67]. They also occur in complex contexts along with many other psychological processes such as orienting, startle, and defense responses, attention, and social interaction [67]. Finally, emotion-relevant ANS activity is superimposed on other physiological activity responsible for contributing to internal processes (e.g., resting and digesting, metabolic needs), and external demands (e.g., orienting, startle, and defense responses) [67]. Although it is not important to identify specific emotions in order to use psychophysiology as an objective evaluation methodology for entertainment technology, these issues are important when interpreting physiological data gathered from the sensors.

4.5.2.2 Physical Activity

There is no question that physical activities can overwhelm the physiological readings from psychological events [102]. Even in a laboratory under very controlled resting conditions, physiological responses to physical needs have to be accounted for. When evaluating entertainment technology, users will move their hands, arms, and perhaps their whole bodies as they interact with the technology and with each other. Acknowledging the effects of physical activity is important, as many physiological sensors produce movement artifacts and physical movement affects many of the physiological measures.

4.5.2.3 Inferring Significance

Aside from the methodological issues, there are also theoretical challenges associated with inferring psychological significance from physiological data. It is imperative to acknowledge the theoretical limitations to ensure a valid experimental methodology. After having identified correlations between events related to the task, psychological events, and physiological data, the eventual goal of an evaluation methodology is to be able to index psychological events from sensor readings. Although possible, there are many issues to address. For a very simple example, if when playing a computer hockey game, a user's GSR reading drops after every period of a hockey game and rises at the beginning of the next period, it is apparent that arousal is lower between hockey periods. However, basic logic prevents us from thinking that a lower arousal between hockey periods means that every time a user's GSR drops, they are in between hockey periods. This seems like an obvious example, but it illustrates the care that must be taken when making inferences.

Cacioppo discusses four classes of psychophysiological relationships called outcomes, markers, concomitants, and invariants [13]. These relations are based on the specificity (one-to-one vs. many-to-one), and generality (context-bound vs. context free) of the relationship between the psychological event and the physiological response (see Figure 11). Outcomes are many-to-one, situation-specific relationships, and reflect the fact that a physiological response varies as a function of a psychological event in a specific situation. When the physiological response follows the psychological event across situations (generality), the relationship is concomitant (many-to-one, cross-situational associations). With outcomes and concomitants, it is

unclear whether the physiological response *only* follows changes for that psychological event or whether other psychological events (specificity) can also inspire the same physiological response. Markers are one-to-one, situation-specific relationships, and reflect that a physiological response can predict the occurrence of a psychological event in a specific situation. Invariants are like markers, except that the psychophysiological relationship is maintained across situations (one-to-one, cross-situational associations). Invariants provide a strong basis for psychological inference. The issue for a researcher is in *establishing* the invariant relationship instead of simply *assuming* that the relationship between a psychological event and a physiological response is an invariant.

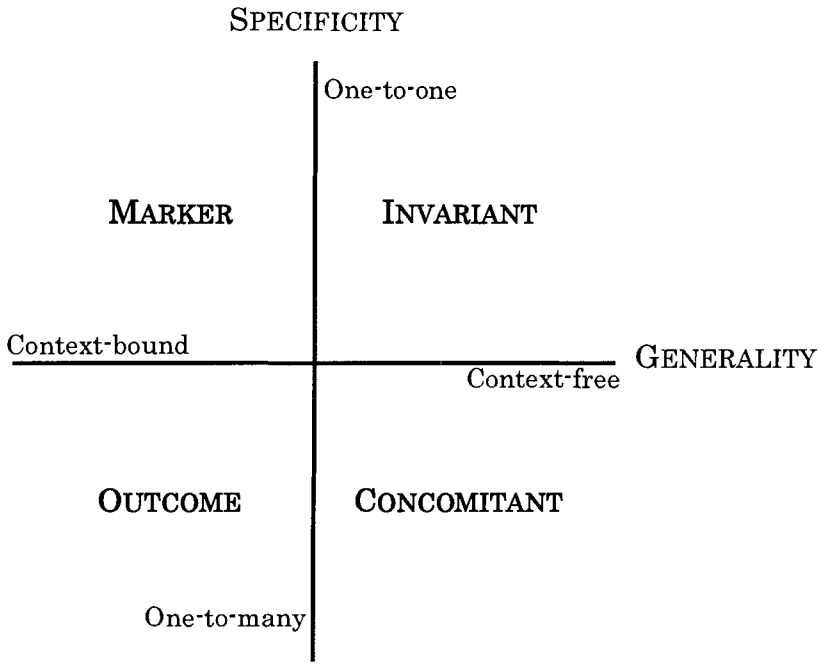


Figure 11: The four types of relationships between psychological events and physiological responses as defined by Cacioppo [13].

Chapter 5 RELATED LITERATURE ON PSYCHOPHYSIOLOGY AS A METRIC FOR THE EVALUATION OF INTERACTIVE SYSTEMS

Traditional evaluation methodologies are presented in Chapter 2, and background information on psychophysiological techniques are presented in Chapter 4. This chapter discusses the current state of using psychophysiological techniques for the evaluation of interactive systems.

The field of human factors has been concerned with optimizing the relationship between humans and their technological systems. The quality of a computer system has been judged not only on how it affects user performance in terms of productivity and efficiency, but on what kind of effect it has on the well being of the user. Psychophysiology demands that a holistic understanding of human behaviour be formed from the triangulation of three fundamental dimensions: overt behaviour, physiology, and subjective experience [141].

Few investigations have been conducted to determine whether psychophysiological techniques could be of use when evaluating HCI issues in computer software. Most of the previous experiments have been concerned with identifying stress and high mental workload in typical productivity environments such as web page navigation [139],

videoconferencing [148], and air traffic control [113]. In this chapter, we first present examples of the use of psychophysiological measures in laboratory tasks, followed by examples of psychophysiological measures used in field studies. There has been an abundance of work conducted in the traditional Human Factors domains of study of dispatch, air traffic control, and simulators. Although these experiments do not deal directly with the evaluation of entertainment, they are the best indicators of the use of psychophysiological measures to evaluate interactive systems. Of most interest is similar research in the domain of HCI; however, this area is where the least amount of work has been conducted.

5.1 Laboratory Tasks

Experimental psychologists and human factors specialists have performed many controlled laboratory experiments in order to investigate the feasibility of psychophysiological measures as an evaluative tool. These controlled experiments have mostly focused on understanding how we physiologically respond to different stimuli. A few examples follow.

Measuring EKG from the forearms, and respiration using a strain gauge sensor around the chest, Sammer [116] computed HRV and IBI for a mental task (addition of numbers in an array), a physical task (moving a lever), and a combination of both. Sammer found that IBI increased from the dual task, to the physical task, to the cognitive task. In addition, HRV was suppressed for the dual task in both the baroflex-related and the respiration-related frequency bands, but no difference of task factor was found between the cognitive and physical tasks.

Fournier et al. [38] collected subjective ratings, eye blinks, respiration rate, respiration amplitude, HR, and HRV in the baroflex-related (.06-.14Hz) and respiration-related (.15-.40Hz) bands while subjects performed either a single task or three multiple tasks of increasing difficulty. The single task was a communication task, while a monitoring, tracking, and detection task was added to the communication task to create the multi-task. Difficulty of the multi-tasks was manipulated by varying the number of events requiring attention. They found that respiration rate, eye blink rate, duration, and amplitude, HR, and HRV in both bands differentiated between the single tasks and the multi-task workload. Within the three multi-tasks, they found only that HR could differentiate the high difficulty task from the mid and low difficulty tasks, and that HRV in the baroflex related band could differentiate between the low and high difficulty tasks. This corresponded with the subjective scores for workload, which were higher in the multi-task medium and high workload conditions than the multi-task low workload condition.

Boutcher, Nugent, McClaren and Weltman [11] collected HRV while participants performed either an arithmetic task or a Stroop task [124]. There was no difference in measured HRV between rest and the arithmetic task; however, HRV decreased for both mid (.07-.11 Hz) and high (.12-.40 Hz) bands when performing the Stroop task, versus resting.

5.2 Field Tasks

Although the following field studies are not directly related to measuring entertainment using psychophysiological measures, they have some similarities.

Laboratory tasks are nicely controlled, but one must measure entertainment in context, increasing the complexity of an experimental setting. Physical movements, changing graphics, and the effects of narrative and plot all have impact on the experimental protocol. In addition, the relationship between subjective measures and psychophysiological measures is very important. The following three examples illustrate some of these issues and how they were handled.

The majority of airplane pilot and air traffic control researchers rely on cardiovascular measures of heart rate, interbeat interval, and heart rate variability, along with brain activity through EEG, eye blinks, and subjective measures. Wilson gathered physiological measures during actual flights from 10 male pilots flying a Piper Arrow [145]. The study also examined electrodermal activity (GSR), and EMG of the leg. Electrodermal activity was gathered on the sole of the foot, thus the EMG was gathered to determine whether leg movements influenced GSR on the foot. Wilson found that takeoffs and landings produced the greatest psychophysiological changes, especially in terms of increases in HR and GSR. They believe that the increased cognitive demands during these flight sequences are highlighted by the physiological changes. In addition, examination of leg EMG revealed that leg movements did not influence the electrodermal responses.

Richter et al. [110] conducted a field study that considered subjective measures and multiple physiological measures. They collected subjective difficulty, electrodermal response, blink rate, heart rate, and heart rate variability while 31 student drivers drove real cars on rural roads with varying rates of curvature change. They found that

subjectively perceived difficulty of the road varied with curvature change rate and that blink rate consistently decreased with increases in curvature change rate of the road. In addition, the number of spontaneous fluctuations of GSR increased with curvature change rate. They also found that HRV was highest during baseline measurements, and then continuously decreased as curvature change rate rose, until the highest curvature change rate, when HRV increased again. For the highest curvature change rate, the increase in HRV could be due in part to a decrease in speed. It could also be due to overwhelming workload.

Healey and Picard [47] used EKG, GSR, Respiration, and EMG to detect driver stress for ten drivers on real roads. They created four stress categories from the participants' subjective ratings and found that they could recognize driver stress at a rate of 88.6% using combinations of the physiological signals.

Myrtek et al. [88] measured HR, HRV (using an ambulatory monitoring device), and subjective stress levels of 50 female university students throughout their day. They found that although physical activity was higher during leisure, HR was higher during study time. The cognitive aspects of study overwhelmed the effects of physical activity on HR. The students rated leisure activities as more enjoyable, but less arousing or exciting than studying, and HRV was suppressed during university-related activities (e.g., studying). In addition, chronically stressed students (determined by a pre-test) had higher HR values, and lower HRV, indicating greater mental workload. No correlations between subjective variables and physiological variables were attempted.

5.3 Dispatch, Air Traffic Control, and Simulators

There has been an abundance of work conducted in the traditional Human Factors domains of dispatch, air traffic control, and simulators. These experiments are generally concerned with the evaluation of cognitive workload and mental effort. Although not directly related to the evaluation of fun and engagement, it is the most well studied domain for the use of psychophysiological measures to evaluate interactive systems. Although there is a large amount of work in this area, we only present examples of some of the seminal research conducted using a variety of tasks.

Wastell and Newman [141] used the physiological measures of blood pressure and heart rate in conjunction with task performance and subjective measures (Likert scales) to determine the stress of ambulance dispatchers in Britain as a result of switching from a paper-based to a computer-based system. When normalized for job workflow, systolic reactivity showed that dispatchers stress increased more for increases in workload in the paper-based system than in the computer system. This was consistent with non-significant results obtained from the post-implementation questionnaires. The authors concluded that this triangulation of data sources provides compelling and complementary insight into stress in the work environment, and is more sensitive than subjective ratings or task performance alone.

Using complex tasks in a flight simulator, Veltman and Gaillard [134] measured heart rate, blood pressure, respiration and eye blinks to try to index mental workload. The flight was the primary task, but users were also required to listen to letters of the Dutch alphabet that were presented through headphones and to press a button to

indicate when one of four target letters was presented. They found that respiration was slower and deeper just after landing the aircraft. In addition, they found that power in the mid (.07-.14Hz) and high bands (.15-.50Hz) of heart rate variability was higher just after landing the aircraft than when flying, and that BVP increased just after landing. The only differences found for the presence of the secondary task versus flying alone were that IBI decreased and BP increased. Similar to other field studies [56, 144], the authors found that some psychophysiological measures (e.g., HRV) can differentiate between rest and task, but not well between different tasks, or different levels of the same task.

In another experiment, the same authors [135] studied twelve pilots flying through tunnels with varying levels of difficulty. During the tunnel flying tasks, they also had to perform a memory task with four levels of difficulty which were matched to the difficulty of the flying tasks. Additionally, they flew a pursuit task in between each tunnel task. Subjective ratings of difficulty, IBI, HRV, BP, respiration, and eye blinks were collected. All collected measures discriminated between rest and the tasks, and in this experiment, the measures also differentiated between the tunnel and pursuit tasks. Subjects perceived the tunnel tasks as requiring more effort than the pursuit task, and also rated the more difficult tasks as requiring more effort. For discriminating between the different difficulty levels of the tunnel tasks, only IBI revealed consistent results. IBI systematically decreased with increasing levels of task difficulty.

Rau [109] collected HR and BP data during simulations of an electrical distribution system with trained operators. Fifty operators were tested in pairs, with one operator acting as the leader, and the other as the co-operator. Three tasks chosen to reflect different levels of cognitive workload were performed. The operators also gave subjective responses for strain, emotion, motivation, perceived control, and success by means of a pocket computer. For the task factor, HR was lower for the least demanding condition than the two higher demand conditions, while perceived strain and mental effort were higher for the demanding conditions. The operators acting as leader showed higher HR, systolic BP, and reported strain than the co-operators, while the co-operators perceived more control and success than the shift leaders for all three tasks.

Cnossen et al. [17] collected HRV in a driving simulator, while subjects drove either fast or accurately (resulting in a lower speed). They repeated each condition twice and were required to attend to a secondary memory task in half of the trials. They found that neither speed nor HRV were influenced by the presence of the secondary memory task, although they rated the memory task as more demanding and their mean HR was higher during the memory task for both fast and accurate driving.

In an air traffic control (ATC)-inspired task, Rowe et al. found that users with ATC experience showed a significant decrease in HRV when the complexity of the task was increased, but found that HRV increased when the complexity increased beyond a threshold value [113]. This increase may be an indication that the task became too

difficult and users weren't trying as hard to achieve good results or that the demands of the task exceeded the capacity of working memory [113].

The aforementioned studies collected both subjective measures and psychophysiological data. Although the authors reported patterns in both of these types of measures, they did not attempt to determine whether the subjective and objective responses co-varied. This is due in part to the difficulties of dealing with high individual variability. Using a hovercraft simulation display, Vicente et al. collected HRV. Instead of examining the raw scores, they normalized HRV to a ratio between 0 and 1 by dividing the HRV scores for each task by the average HRV scores of the rest periods. They were then able to correlate subjective ratings to physiological data. They determined that normalized HRV significantly correlated to subjective ratings of effort, but not workload or task difficulty [137]. Participants were instructed to rate effort as the amount of attentional demand they allocated to the task, or how hard they were trying, while workload was rated on the overall level of demand imposed by the task, and difficulty was rated on how hard the task was.

5.4 Adaptive Technologies

Adaptive technologies seek to trigger changes among modes based on real-time performance, critical events, or operator models [118]. Recently, researchers in this area have become interested in the use of psychophysiological signals to reflect changes in operator workload. Some psychophysiological measures have the advantage that they can be obtained continuously, in real-time, with little computation. In addition, although physiological sensors can be very clinical, and personally

invasive, they do not interfere with an operator's task in the same way as a think-aloud protocol, or a randomly-prompted Likert scale [118].

Wilson, Lambert, and Russell [146] monitored EEG and respiration while subjects performed tasks in two difficulty levels, and a resting condition. Physiological data from the sessions were used to train a neural network to recognize these three different conditions. The goal was to demonstrate that performance on these tasks could be improved by adaptively reducing the number of subtasks when high levels of mental workload were detected. This adaptive technique, based on physiological signals, reduced errors in a tracking task by 44% and in a monitoring task by 33%. Although successful, these results reflect that participants exhibited improved performance when mental workload was decreased, regardless of adaptation.

Piechulla et al. [103] estimated workload for drivers in order to create an adaptive interface. Dividing attention between driving and talking on cell phones can cause traffic accidents, even when the cell phone is hands-free. In the authors' system, whenever workload estimates (from simulation data) exceeded a threshold, incoming calls were automatically redirected to voice mail without notifying the driver. Both subjective measures and psychophysiological estimates of workload were used to favourably assess the system in a field study. Piechulla et al. used HRV and Lateral Frontalis EMG (eyebrow-raising muscle) as estimates of workload.

Rani et al. [108] used HRV to determine the stress of an online robot operator in order to create a robot that could respond accordingly. They classified stress using a fuzzy logic model. Using sympathetic and parasympathetic heart activity through HRV

calculations, they modeled the stress level of two participants in an experimental situation that involved playing video games. No comparison to another data source (such as subjective ratings) were made, and the experimental session only involved playing the game in one session, so no comparisons were made across difficulty levels. In a subsequent study, Rani et al. [107] collected HRV, GSR, and EMG of the forehead and jaw, for one participant who performed problem solving tasks. A fuzzy logic model transformed the physiological variables into an anxiety index, which exhibited the same trends as the subject's self-reported anxiety levels.

5.5 Psychophysiology in Human-Computer Interaction

Wilson (and Sasse) [147-149] employed the triangulation of data sources to evaluate subject responses to audio and video degradations in videoconferencing software. Describing their approach as three-tiered, the authors suggest that subjective ratings of user satisfaction and objective measures of task performance be augmented with physiological measures of user cost (impact of media quality on the user) [147]. Using three physiological signals to determine user cost, they found significant increases in GSR and HR, and significant decreases in BVP for video shown at 5 frames per second versus 25 frames per second [148], even though 84% of subjects did not report noticing a difference in media quality. In another experiment, significant physiological responses (increase in HR, decrease in BVP) were found for poor audio quality [149], but these results weren't always consistent with subjective responses. For example, a bad microphone induced more physiological stress than a quiet speaker or a 5% packet loss, but was only rated subjectively worse than the 5% packet loss. Also, an induced echo was physiologically more stressful than a quiet

mike, but was not rated worse by the participants. These discrepancies between physiological and subjective assessment must be noted. In this audio experiment, a bad microphone was the first and second most stressful condition physiologically but was not subjectively rated as poor. If only subjective ratings were considered, the effects of a bad mike on quality of media would have been missed.

Ward et al. [139, 140] collected GSR, BVP, and HR while subjects attempted to answer questions by navigating through either well- and ill- designed web pages. No significant differences in physiological measures were found for navigating the two types of web pages, which is not surprising considering the large individual differences associated with physiological data, and the between-subjects experiment design. However, distinct trends were seen between the two groups when the data were normalized and plotted. Users of the well-designed website tended to relax after the first minute whereas users of the ill-designed website showed a high level of stress for most of the experiment (exhibited through increasing skin conductance and heart rate). Because the data were collected in a naturalistic setting, using a real-life task, rather than with pure stimuli in a controlled environment, the data show promise for using physiological data to evaluate HCI issues. Using trends in skin conductivity alone, the authors suggest that it may be possible to distinguish between low, medium, and high levels of stress in the user, both over periods of time and as a direct result of an event in the interface. It is important to note that these authors achieved only somewhat reliable results when distinguishing between two well-understood, already evaluated web site designs. In many HCI evaluations, there isn't a controlled

environment with baseline performance that evaluators can compare their design to, making the task of recognizing and interpreting physiological data more difficult.

Partala and Surakka [98] used facial EMG activity to investigate affective audio intervention during computer malfunction. Audio intervention informed the user that the system was not functioning properly and then provided a short emotional statement using either positive or negative terms (e.g., 'great that it will be working again soon' versus 'sorry this is so frustrating'). They recorded EMG activity on the zygomaticus major (smiling muscle) and corrugator supercilii (frowning muscle) while participants performed a problem-solving task on a computer with preprogrammed mouse delays. Following the task, positive, negative, or no audio intervention regarding the mouse delay was provided. Smiling activity was higher during positive feedback than during negative feedback, and after either intervention, smiling activity was higher than after no intervention. Frowning activity attenuated significantly more after the positive intervention than no intervention. In addition, performance improved more following the positive intervention than no intervention.

Chen and Vertegaal [16] used brainwave activity and HRV to distinguish between four attentive states of a user: at rest, moving, thinking, and busy. They used this approach to change the interruption behaviour of a cell phone, and found during a six-person trial, that the system identified the appropriate notification level in 83% of the trials.

There has not been much attention paid to using psychophysiology to objectively evaluate entertainment technology. Sykes and Brown [128] measured the pressure

that gamers exerted on the gamepad controls and correlated this with game difficulty; however, there has not been much work on using psychophysiology to create direct metrics of human experience with entertainment technology.

5.6 Affective Computing

Affective computing [101] is described as "computing that relates to, arises from, or deliberately influences emotions" [101]. Research in affective computing is concerned with having computers respond to our emotional state and introducing emotional responses into our computers. Potential domains of use for affective technologies span work applications, travel, communication, and entertainment.

For example, a few research groups have been interested in creating affective cars that can analyze the stress of the drivers [47, 48, 89]. This stress level information could be used to automatically adjust non-critical systems such as music selection and climate control. Gathered over time, records of stress-induced alterations provide a valuable source of data to the driver as well as a holistic representation of driver stress.

5.6.1 Psychophysiology as an Input Stream

Physiological measures have traditionally been used for evaluating stress and for biofeedback applications. For affective computing researchers, there is a new interest in using physiological data as an input mechanism, instead of or in addition to explicit input through mice, keyboards, and other controllers. New sensing technologies facilitate this interest, in which some input devices will evolve from current explicit

manipulation of electromechanical devices to the implicit input of subtle human physiology [3].

For example, physiological data has been used as an input stream for communications technologies. In Conductive Chat [25], an instant messaging client incorporates users' fluctuating skin conductivity levels into the dialogue interface. The size and color of the font are adjusted based on the skin conductance of the user. In this way, collaborators have a visual display of their partner's level of arousal without any explicit input.

Replacing the input devices of mobile technologies with smaller, more context aware technologies is another arena for physiological input. The Biofeedback Pointer is a graphic input device that operates by sensing EMG signals of the wrist and interpreting this data with a neural network to determine where the user is pointing [112]. Using Fitts's law [36, 37], the index of performance (a measure of the information capacity of the human motor system) of this prototype device was found to be only 1.06 as compared to the mouse at 7.10 [112]. Isometric EMG of the arm has also been used for subtle and intimate input, requiring very little movement from the user. Without calibration or training, an armband EMG device was able to reliably recognize a motionless gesture across users with different muscle volumes [19].

Another line of research uses gaze for targeting and voluntary facial muscle activation (from EMG readings) for selection [125]. The eye is a great targeting tool, but as it is a perceptual organ, gaze as a selection mechanism suffers from accidental activation [152]. Surakka et al. [125] implemented a system that used gaze for targeting and

corrugator supercilli (eye frowning) activation for target selection. They compared their method to mouse selection and found that the mouse was faster for close targets, but that there was no difference between the mouse and their new technique for mid and far targets. In fact, their regression analysis showed that at very far distances, their gaze and frown technique might be superior to mouse control. Their gaze and frown technique was also much more prone to errors than mouse control. Developments such as the biofeedback pointer and the gaze and frown technique provide good starting points for developing physiological input devices.

There is great potential for enriching entertainment technology with physiological input. Current entertainment applications using physiological input include a music selection DJ that picks music based on a user's affective state [46], and a jacket worn by a conductor that can create music and visualizations of music based on the conductor's affective state [78]. There are also a few examples of using affective state as an input to a game environment. AffQuake [2] alters game play in the popular Quake first person shooter game with GSR signals from a player. For example, when a player is startled, the player's avatar is also startled and jumps back. AffQuake also relates the size of the player's avatar to the arousal of the player. Brainball [50] is a game where brain waves (from EEG) are used to alter the direction that a physical ball rolls on a physical table. Players sit across from each other and need to relax to make the ball move towards their opponent.

There also have been a few games developed as biofeedback applications to treat disorders such as stress and Attention Deficit Hyperactivity Disorder (ADHD). In

Relax To Win [7], a player controls the speed of a racing dragon through GSR. As a player relaxes, their dragon moves faster. This was also the principle behind a commercially unsuccessful car racing game released by Human Engineered Software and promoted by Leonard Nimoy. Using NASA technology, S.M.A.R.T. Braingames [115] uses real Playstation video games as a biofeedback application. Using EEG signals, the system determines whether the user is in the desired brain state, and adjusts accordingly. If the user does maintain the desired brain state, full control of the game controller is provided. If not, the speed and steering control decreases. Basically, as the player maintains their focus, the game responds, and when they lose their focus, they lose ground. Researchers tested this game environment against a traditional biofeedback training environment and found no difference between the successes of the two systems in terms of clinical improvements, but found that both parents and children preferred using the video game system [62].

In the Affective Computing Lab at MIT, researchers have sensed a game player's affective state while playing DOOM, but have not released any information on how they propose to use this data [101]. An extensive literature search has not revealed any use of physiological data to create direct metrics of human experience (e.g., for use as a tool for game developers), or to deepen engagement with a console or computer game as an input.

5.6.2 Wearable Biometrics

Because of the clinical and personally invasive nature of physiological sensors, many research groups are creating intimate, wearable sensors [102]. Commercial initiatives

for smart fabrics are abundant. For example, Philips created smart underwear (bras and briefs) that monitor your heart rate and dial for help in case of an emergency [100]. Based on research conducted at the Georgia Institute of Technology [44], the Sensatex smart shirt collects biometric information such as heart rate, respiration rate, body temperature, and caloric burn, and provides readouts via a wristwatch, PDA, or voice. Biometric information is also wirelessly transmitted to a personal computer and ultimately, the Internet [121]. The LifeShirt garment collects over 30 physiological signals related to pulmonary cardiac and posture data [69]. Optional peripherals can collect many other physiological states including blood pressure, blood oxygen saturation, EEG, periodic leg movement, core body temperature, skin temperature, and cough.

Chapter 6 EXPERIMENT ONE: GOLDILOCKS

To begin to understand how physiology can be used to objectively measure user experience with entertainment technology, we collected a variety of physiological measures while observing participants playing a computer game. Participants played in four different conditions of difficulty: beginner, easy, medium, and difficult. We called this initial experiment Goldilocks because of these game difficulty manipulations. Our goal was to either create an experience that was too easy; that was too hard; or that matched a player's experience to the difficulty level in the game, creating a condition that was 'just right'.

We expected that participants would prefer playing in the condition that was best matched to their level of expertise, experiencing the most enjoyment, satisfaction, and engrossment in this condition. These preferences would be reflected in their subjective experience as well as their physiological experience. Our previous studies on play technologies, as well as the literature on physiology and emotion were used to generate the following experimental hypotheses.

H1: *GSR will increase in conditions where players report a greater sense of fun and excitement, and a lesser sense of boredom.*

H2: *EMG of the jaw will increase in conditions where players report a greater sense of challenge and frustration.*

H3: *Respiration Rate will increase in conditions where the players experience greater challenge.*

6.1 Participants

Eight male participants were recruited from computer science and engineering students at Simon Fraser University to participate in the experiment. We chose to test only male participants in order to reduce any potential confounds since females respond differently to computer game environments, and also have different physiological and emotional reactions in general.

Participants were given a free game from EA Sports to thank them for their participation in the experiment. One participant did not complete the experiment, so we have data for seven participants aged 20 to 26[6]. Before participating in the experiment, all participants filled out a background questionnaire (see Appendix 5). The questionnaire was used to gather information on their computer use, experience with computer and video games, game preference, console exposure, and personal statistics such as age and handedness.

---

[6] An eighth participant was recruited to balance the experimental design; however, the laboratory where the experiment was conducted was dismantled before we were able to test the final participant. Testing the eighth participant under different laboratory conditions would have introduced confounds and issues with the analysis and interpretation of the data.

All participants were frequent computer users. When asked to rate how often they used computers on a 5-point Likert scale (1-5), all seven subjects used them every day (corresponding to 5). The participants were also all self-declared gamers. When asked how often they played computer games, two of the participants played every day, four played often, and one played occasionally. For the frequencies of responses to questions on computer usage and game play, see Table 1 and Table 2. When asked how much they liked different game genres, role-playing was the favorite, followed by strategy and action games (see Table 3).

Table 1: Frequency of computer usage and game play for Experiment One. Participants were asked to respond to how often they do each of the following activities:

	Never	Rarely	Occasionally	Often	Every day
Use computers?					7
Play computer games?			1	4	2
Play video (console) games?		1	1	3	2
Play computer/video games over the internet or network?			2	3	2
Play computer/video games with another co-located player?		2	1	4	

Table 2: **Frequency of computer usage and game play for Experiment One. Participants were asked to respond to how much time they spend doing each of the following activities:**

	Never	< 3 hours a week	3-7 hours a week	1-2 hours a day	> 2 hours a day
Use computers?					7
Play computer games?		1	3	2	1
Play video (console) games?		1	3	2	1
Play computer/video games over the internet or network?		2	3	1	1
Play computer/video games with another co-located player?		3	3	1	

Table 3: **Results of game genre preference from background questionnaires for Experiment One. A 5-point Likert scale was used with "1" corresponding to "Dislike a lot" and "5" corresponding to "Like a lot".**

	Mean	St.Dev.
Action	4.57	.54
Adventure	4.29	.76
Puzzle	3.00	1.2
Racing	3.57	1.4
Roleplaying	4.86	.38
Shooting	4.57	.79
Simulation	3.86	1.1
Sports	3.86	1.2
Strategy	4.57	.54

6.2 Play Conditions

Participants played the game in four difficulty conditions: beginner, easy, medium, and difficult. To balance the order of presentation of the difficulty conditions, we used a reversed Latin Square design. The order of the conditions was either BEMD, EMDB, MDBE, DBEM, or the reversed DMEB, MEBD, EBDM, BDME, where B stands for beginner, E for easy, M for medium, and D for difficult.

Participants played NHL 2003 by EA Sports in all conditions (see Figure 12 for a screen shot). In the background questionnaire, we asked participants to state how experienced they were with NHL 2003 or previous versions of the game. When asked to rate their experience on a 5-point scale, three of the participants selected "very experienced", one selected "somewhat experienced", two chose "somewhat inexperienced", and one chose "very inexperienced". As a result, we had three players who were experts, three players who were novices, and one player who had played the game in the past, but did not consider himself an expert.

Each play condition consisted of one 5-minute period of hockey. The game settings were kept consistent during the course of the experiment. All players played the Dallas Stars while the computer played the Philadelphia Flyers. These two teams were chosen because they were comparable in the 2003 version of the game. All players used the overhead camera angle, and the home and away teams were kept consistent. This was to ensure that any differences observed within subjects could be attributed to the change in play setting, and not to the change in game settings, camera angle, or direction of play.

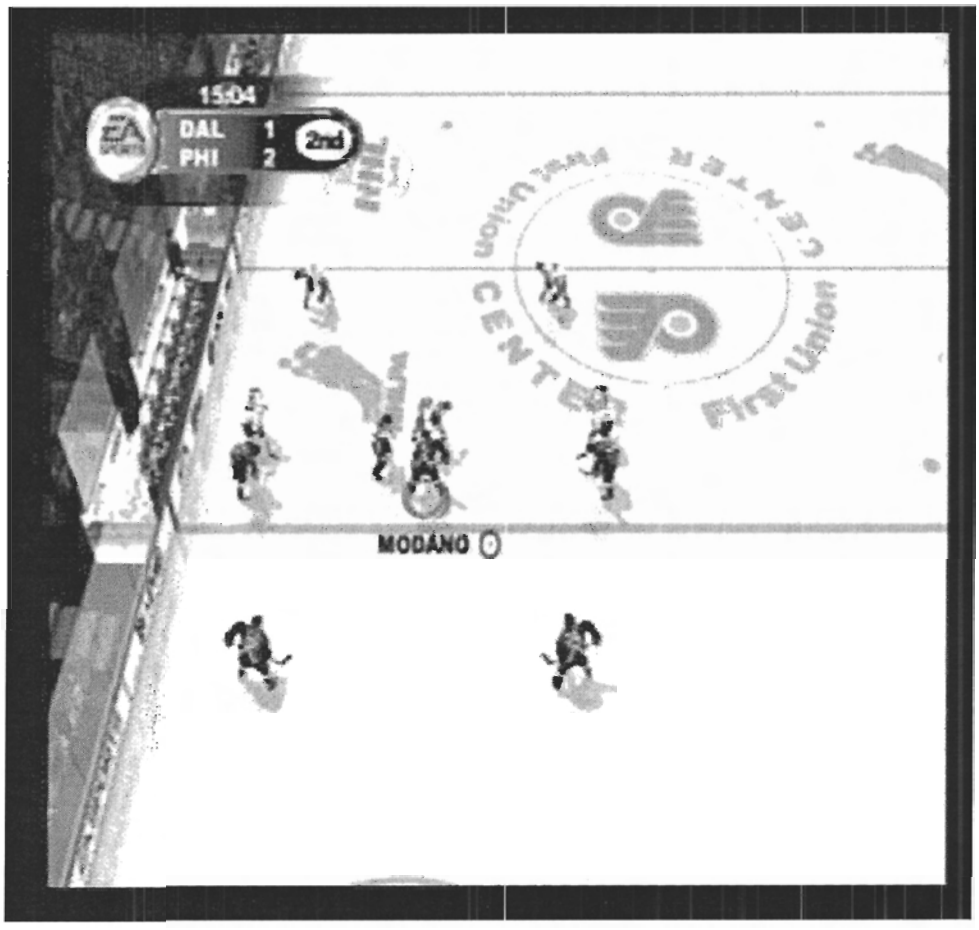


Figure 12: Screen shot of NHL 2003 by EA Sports.

6.3 Experimental Setting and Protocol

The experiment was conducted in the Human Centered Design Laboratory at the New Media Innovation Centre, located in downtown Vancouver. NHL 2003 was played on a Sony PS2, and viewed on a 36" television. Cameras captured a player's facial expressions and their use of the controller. All audio was captured with a boundary microphone. The game output, the camera recordings, and the screen containing the physiological data were synchronized into a single quadrant video display, and recorded onto a hard disk (see Figure 13). The audio from the boundary microphone, and the audio from the game were integrated into the exported video file.

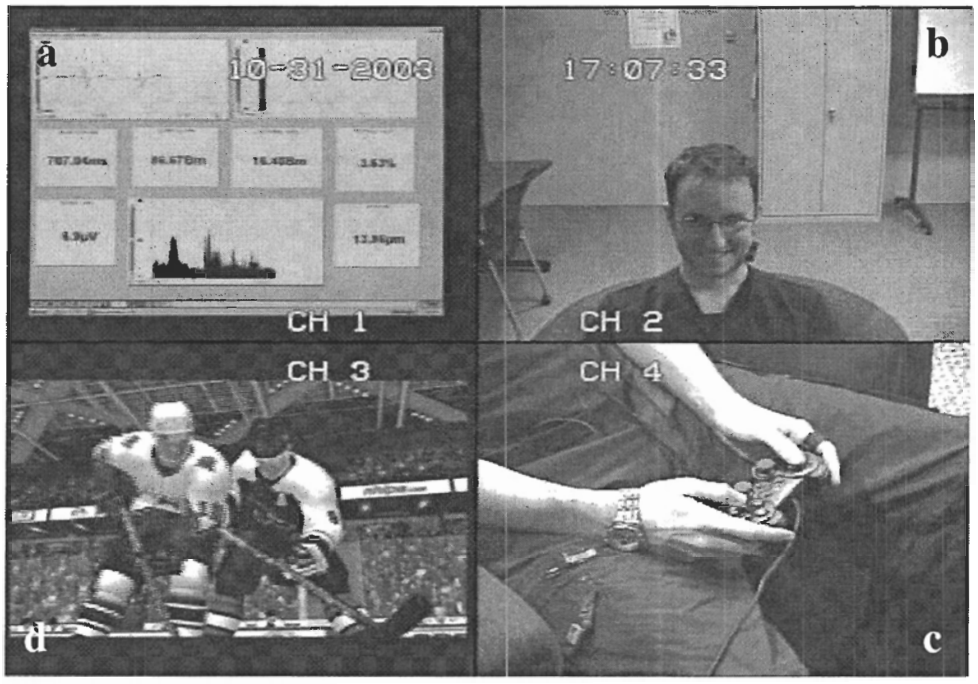


Figure 13: Quadrant display for Goldilocks including a) the screen capture of the biometrics, b) video of the participant's face, c) video of the controller, and d) a screen capture of the game. Audio of the participant's comments and audio from the game were included in the quadrant video.

Physiological data were gathered using the ProComp Infiniti system and sensors (see Figure 14), and BioGraph Software from Thought Technologies. Based on previous literature, we chose to collect galvanic skin response (GSR), electrocardiography (EKG), electromyography of the jaw (EMG_{jaw}), and respiration. Heart rate (HR) was computed from the EKG signal, while respiration amplitude (RespAmp) and respiration rate (RespRate) were computed from the raw respiration data. We did not collect blood volume pulse data (BVP) because the sensing technology used on the finger is extremely sensitive to movement. As our subjects were operating a game controller, it wasn't possible to constrain their movements.

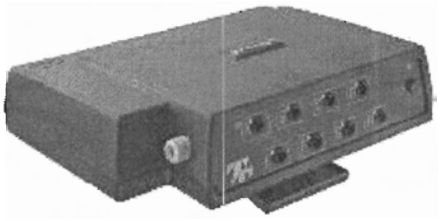


Figure 14: The ProComp Infiniti system from Thought Technologies. Sensors described in Chapter 4 were connected to the 8 ports in the front of the unit, while the unit was connected to the computer's serial port via fiber-optic cable and a serial port converter.

Upon arriving, participants signed a consent form (see Appendix 4), after which they were fitted with the physiological sensors. The participants then rested for 5 minutes, after which they played the game in their first difficulty level. After each difficulty condition, the primary experimenter interviewed the participants, using the questionnaire in Appendix 6. Participants were asked to rate the challenge, frustration, boredom, and fun of each condition on a 5-point scale ("1"=low, "5"=high). Explanation of their answers was encouraged. After completing the experiment, the same experimenter interviewed the participants again, using the questionnaire in Appendix 7. Participants were asked to rank the four difficulty conditions in terms of challenge, excitement, and fun. Again, they were encouraged to explain their answers.

6.4 Data Analyses

The subjective data from both the condition questionnaires and the post experiment questionnaires were collected into a database, and analyzed using non-parametric statistical techniques.

In terms of the physiological data, EKG data were collected at 256 Hz, while GSR, respiration, and EMG_{jaw} were collected at 32 Hz. HR, RespRate, and RespAmp were computed at 4 Hz. Physiological data for the rest period and each condition were exported into a file.

When the ProComp Infiniti system records information from the EKG sensor, it may, on occasion, record extra information that is not related to the EKG signal [129]. This could be very small electrical activity from a nearby muscle group or some electrostatic noise that is picked up from the environment. Whatever the source, this noise may confuse the software and cause it to erroneously calculate HR values in two different ways. A sudden surge in the recorded voltage may be interpreted as an extra heartbeat, or a real heartbeat may be lost in the noise, causing the BioGraph software to miss a beat. Using a method prescribed by Thought Technologies, producers of the ProComp Infiniti and Biograph software [129], we corrected these erroneous computations by inspecting each HR sample and contextualizing it in the surrounding samples. If the value of sample was clearly half the value or double the value of surrounding samples, it was corrected. For each condition and the rest period, HR data were then computed into the following measures for each participant: mean HR, peak HR, min HR, and standard deviation of HR. The same four measures (mean, peak, min, and standard deviation) were also computed on the GSR data, EMG_{jaw} data, RespAmp data, and RespRate data. We did not compute HRV. The computation involves a standard-sized time window, and a controlled setting. Due to our ecological approach, we could not ensure that the conditions necessary for HRV analysis were met.

6.5 Results and Discussion

Results of the subjective data analyses are described first, followed by results of the physiological data analyses.

6.5.1 Subjective Responses

Participants rated their experience playing the game in terms of boredom, challenge, frustration, and fun on a 5-point scale after playing in each of the conditions. The mean results are shown in Table 4 and Figure 15. When averaged across participants, boredom decreased, challenge increased, and frustration increased with increasing difficulty in the game. These differences between conditions are a result of averaging across all players; however, when each player is examined individually, there aren't consistent trends. Each player did not have the same subjective experience.

A Friedman test revealed that only the mean ratings for challenge were significantly different across difficulty conditions (χ^2=13.1, p=.004). Although mean perceived challenge increased with every increase in difficulty level, post-hoc analysis revealed that only the beginner level was perceived as significantly less challenging than the medium level and difficult levels (see Table 5). A larger number of participants might yield results where each successive difficulty level is perceived as more challenging than the previous level.

Table 4: Mean subjective responses for each difficulty level. A response of "1" corresponded to "low" and "5" corresponded to "high". Only the ratings for challenge were significantly different across difficulty conditions.

	Beginner	Easy	Medium	Difficult	χ^2	Sig.
Boredom	1.6	1.3	1.0	0.9	4.4	.220
Challenge	**1.0**	**2.0**	**2.4**	**3.3**	**13.1**	**.004**
Frustration	1.0	1.3	1.4	1.4	1.7	.627
Fun	2.9	2.6	3.0	3.3	3.3	.355

Table 5: Wilcoxon Signed Ranks Test results (Z-scores and p values) for perceived challenge. Only the beginner level was perceived as significantly more challenging than the medium and difficult levels.

	Beginner	Easy	Medium
Easy	1.84 (Z) .066 (p)		
Medium	2.23 (Z) **.026 (p)**	1.13 (Z) .257 (p)	
Difficult	2.21 (Z) **.027 (p)**	1.81 (Z) .071 (p)	1.89 (Z) .059 (p)

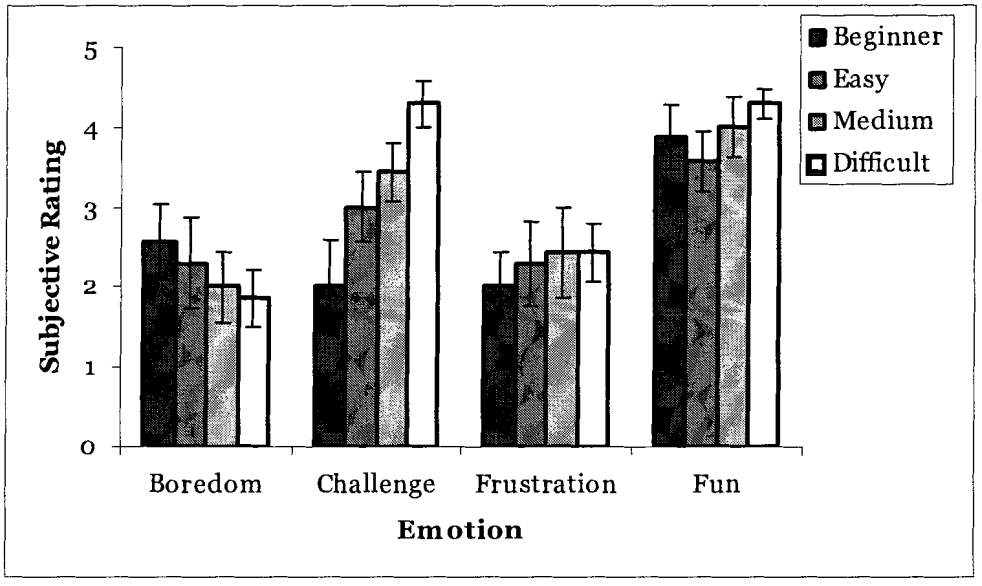


Figure 15: Mean subjective responses (± SE) for each difficulty level. Participants rated four emotional states on a scale from 1 (low) to 5 (high) after each difficulty condition. Difficulty condition only had a significant effect on the challenge ratings.

6.5.2 Physiological Measures

Because the subjective ratings were not consistent across participants, we can infer that the manipulation of the difficulty levels did not produce consistent experiences for all participants. As a result, we did not expect that the physiological results would be consistent across participants. Even so, we used a multivariate analysis of variance (MANOVA) with the four difficulty levels as an independent variable and the three levels of self-identified player expertise as a between-subjects factor, to determine if the level of difficulty or expertise of the player had any measurable effect on the mean physiological measures.

Figure 16 shows plots of the five mean physiological measures, separated by difficulty condition. There were no main effects of difficulty level on any of the physiological measures (HR: $F_{3,12}$ =1.55, p=.252, η^2=.28; GSR: $F_{3,12}$ =.19, p=.899, η^2=.05; EMG_{jaw}: $F_{3,12}$ =1.1, p=.375, η^2=.22; RespRate: $F_{3,12}$ =.78, p=.527, η^2=.16; RespAmp: $F_{3,12}$ =.96, p=.441, η^2=.19). Between subjects, there was an effect of level of expertise on mean respiration rate, measured in breaths/minute ($F_{2,4}$ =24.2, p=.006, η^2=.92). Post-hoc analysis revealed that expert players (mean=33.3, SE=.79) had a higher mean respiration rate than either novice players (mean=27.9, SE=1.4), or semi-experienced players (mean=25.7, SE=.79).

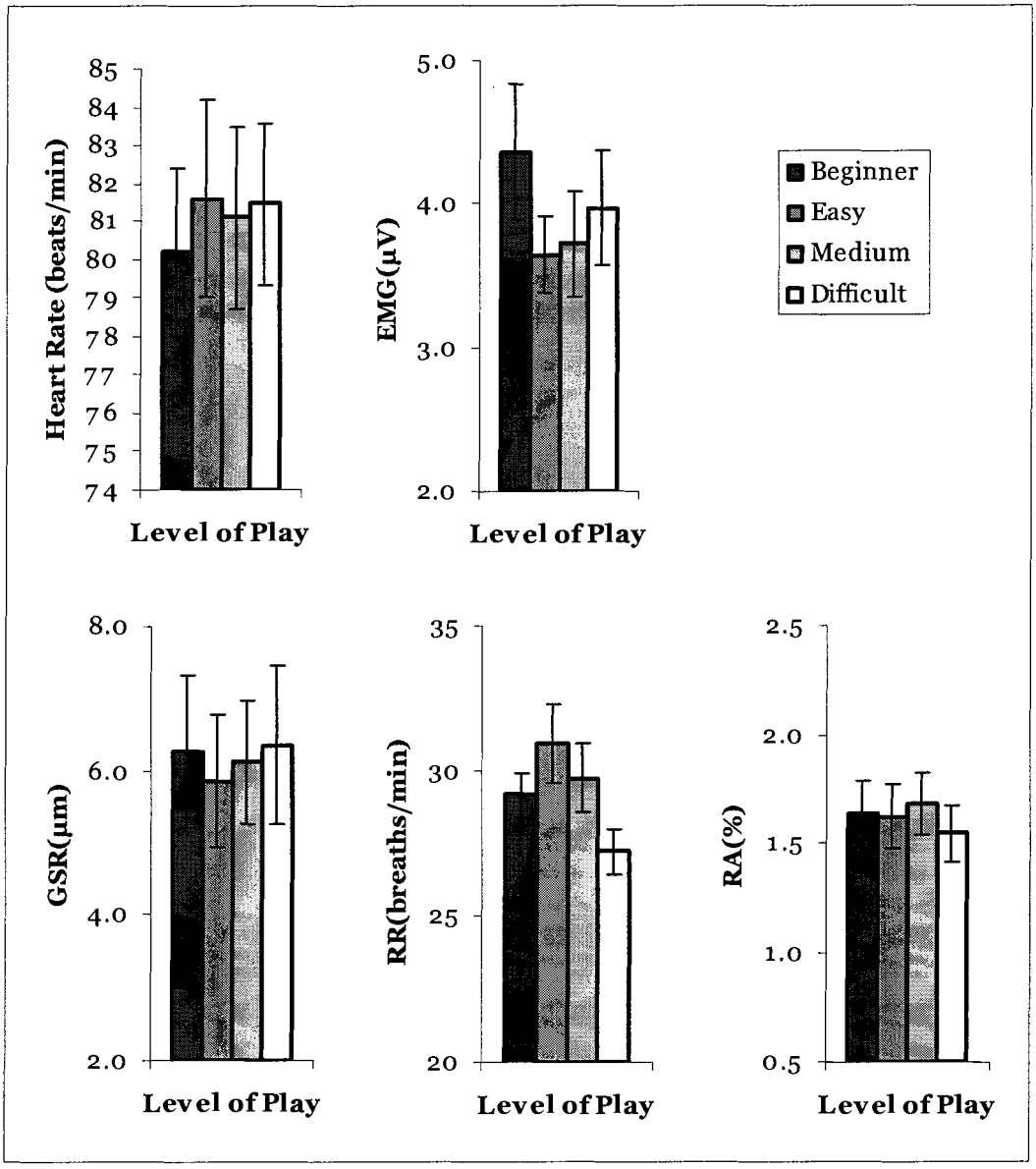


Figure 16: **Mean physiological results (± SE) separated by difficulty condition. There were no main effects of difficulty level on any of the physiological measures. HR: Heart rate; EMG_{jaw}: Electromyography of the jaw; GSR: Galvanic skin response; RR: Respiration Rate; RA: Respiration Amplitude**

As shown in Figure 17, there was also an interaction between difficulty and expertise on heart rate ($F_{6,12}$ =6.03, p=.004, η^2=.75), but not on any of the other physiological measures. The interaction revealed that there was no difference in HR for expert players, but that HR was higher in the easy condition than the beginner, medium, or hard conditions for novice players; and that HR was higher in the difficult condition than the beginner or easy conditions for semi-experienced players. There is no simple explanation for this result, but considering that HR tends to increase with positive affect as compared to negative affect [150], it could be that the game level best matched with the participant's level of expertise produced a positive play experience, generating higher heart rates.

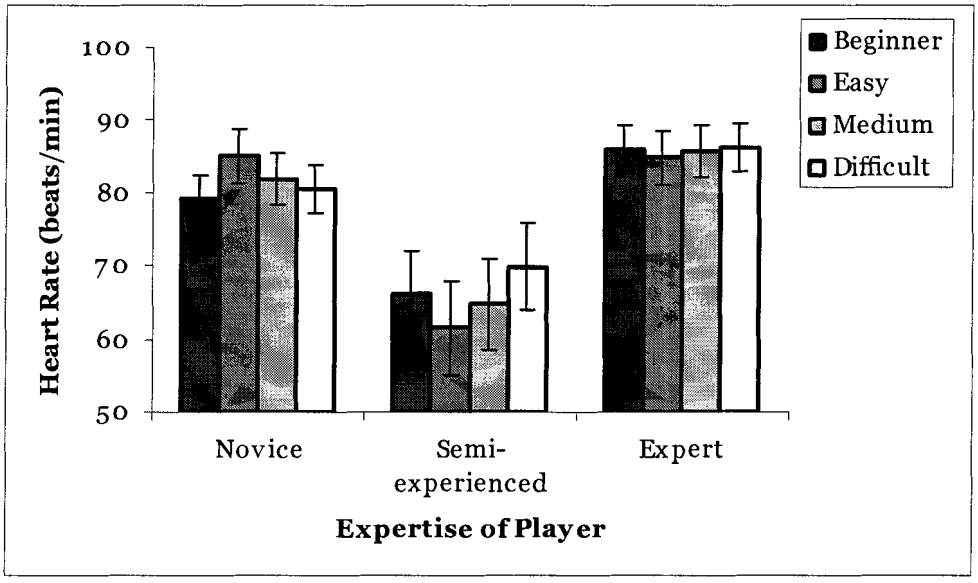


Figure 17: Mean Heart Rate (± SE) split by difficulty condition and expertise. There were no differences between conditions for experts, but there were significant differences for semi-experienced and novice players.

6.5.3 Correlation of Physiological Measures to Subjective Results

Based on our subjective results, we didn't expect that difficulty level would impact the physiological measures, and upon further examination, we discovered that players were not responding consistently to the experimental manipulations.

Although participants did not respond consistently to the difficulty settings, our hypotheses for this experiment expected any given participant's physiological results to correspond to their subjective reports. This doesn't require consistent subjective or physiological responses across participants, just that each individual's physiological responses match with their subjective experience.

Unlike subjective ratings, there are large individual variations in physiological data. We wanted to correlate the subjective ratings to the physiological data, but in order to handle these individual differences we correlated the mean of each physiological measure to the subjective ratings for each participant individually. We then looked to see whether these correlations were consistent across individuals. A relationship between a physiological measure and a subjective rating would be evidenced by a significant number of the participants showing the correlation between the physiological measure and the subjective rating. The individual correlations, and the number of occurrences of each significant correlation are shown in Table 6.

Table 6: Significant correlations between subjective ratings and mean physiological measures for each participant. The four subjective ratings for each of the four difficulty conditions were correlated with the five mean physiological ratings for the four difficulty levels, for each participant. Direction indicates whether the correlation was direct (+) or inverse (-). The number of occurrences represents the number of times the correlation between that subjective rating and physiological measure is seen over all participants. For example, the Challenge-RespAmp correlation is seen three times, (for participants 1, 6, and 7), while the Frustration-HR correlation is seen only once, (for participant 4).

ID	Subjective Rating	Physiological Measure	Direction	Pearson Correlation	Sig.	# occurrences
1	Challenge	RespAmp	+	.967	.033	3
2 *						
3	Boredom	EMG_{jaw}	+	.973	.027	1
	Challenge	GSR	+	.966	.034	2
	Fun	Resp Rate	-	.977	.029	1
4	Boredom	Resp Rate	-	.984	.016	1
	Frustration	HR	-	.977	.023	1
	Frustration	IBI	+	.958	.042	1
	Frustration	GSR	-	.974	.026	1
5	Challenge	GSR	+	.988	.012	2
6	Frustration	RespAmp	-	.950	.050	1
	Challenge	EKG	-	.965	.035	1
	Challenge	RespAmp	+	.997	.003	3
7	Challenge	RespAmp	+	.994	.006	3

* For participant 2, the ratings for boredom, frustration, and challenge were constant. As such, only the ratings for fun were tested, resulting in no significant correlations.

Although there were correlations for most individuals, these correlations weren't consistent across participants. The most common correlation, between challenge and respiration amplitude, only occurred for three of the seven participants. GSR increased with perceived challenge for two of the participants, while all other significant correlations between subjective measures and perceived measures occurred for only one participant. Given the fact that our hypotheses were not confirmed, we needed to determine whether our hypotheses were initially wrong, or whether we were not measuring accurately. Our hypotheses were based on the extensive literature on

physiological responses and emotional states; so in order to explain the inconsistencies between our expectations and our results, we carefully inspected the data. Upon further examination, we discovered that the participants were responding more to the experimental situation than the experimental manipulations. Our methodological decisions were impacting the physiological measures and the subjective ratings in ways we had not anticipated. These issues are discussed further in the next section.

6.6 Issues in Experiment One

There were a number of issues in Experiment One. These issues were mostly methodological, and each is described in detail.

Subjects enjoyed playing in all conditions: One problem was that the subjects enjoyed playing in all of the conditions, even if the difficulty level didn't match their experience. The results of the condition questionnaires showed that the median result for perceived fun was 3.0 for all conditions. Subjects engaged in meta gaming to make the experience more enjoyable, such as by creating challenges for themselves in the easier levels. For example, when playing in the beginner condition, one player set up fancy plays to score pretty goals to make the game interesting since he was able to score at will. Another player tried to get as many goals as possible to see if he could beat his friend who had participated on a previous day. These activities changed the nature of the difficulty conditions, confounding the results. Our pilot subjects had responded to the different difficulty conditions; however, this choice of experimental manipulation did not produce a significantly different experience for the seven subjects in the experiment.

Variability inherent in game play: A significant challenge in analysing this experiment was relating single point data (subjective ratings) to time series data (physiology). To match these two types of data, previous researchers in other domains have converted the time series data to a single point through averaging (e.g., mean) or integrating (e.g., HRV) the time series. This method has been used successfully in the domain of human factors but doesn't apply well to gaming. For example, an air traffic controller would suppress their anxiety and cope with stress, essentially flattening HRV and minimizing variability in other measures. In games, engagement is partially obtained through successful pacing. Variability, in terms of required effort and reward, creates a compelling situation for the player. Collapsing the time series into a single point erases the variance within each condition, causing us to lose valuable information.

High resting baseline: Resting rates were sometimes higher than game play rates for measures where this result is unexpected (e.g., HR, HRV, GSR). Anticipation and nervousness caused the resting baselines to be artificially high. Vicente et al. [137] recommended collecting a number of baselines throughout the experimental session and averaging them to create a single baseline value. In addition, using participants who are familiar with the process of being connected to physiological sensors would help lower the resting values. Beginning the experiment with a training or practice condition, before collecting the resting values, might help the participants to relax. Finally, a relaxation CD used during the resting period may also help to achieve valid resting baselines.

Interview effects: The process of interviewing caused significant physiological reaction from each of the players. This could be because the interviewer was unfamiliar to the participants, of the opposite sex, within their personal space, or simply because the process of answering questions was arousing for the participants. One participant began to stutter during the condition interviews even though he had not stuttered in previous casual conversation with the interviewer. We expect that some combination of these reasons contributed to the participants' reactions.

Order Effects: When examining the data, we noticed that the order of condition may have impacted the results. For example, one participant's GSR signal over the course of the experiment is shown in Figure 18. GSR tends to drift, but note how the increases in the GSR signal over time are catalyzed by the interview. The areas shaded in light grey represent when the participant was being interviewed. The extreme reaction to the interview is seen at the beginning of each light grey shaded area. The areas shaded in dark grey represent when the participant was playing. The GSR signal drops off at the beginning of each game condition from the reaction to the interview process but does not return to baseline levels. These interview peaks cannot be excluded from the analysis, because there were no rest periods in between play conditions. The effects of relaxing post-interview and being excited by the game are inseparable, thus the interview peaks cannot be eliminated.

We cannot include order as a factor in our MANOVA, since we used a Reverse Latin Square design to balance the order of presentation of difficulty conditions. Thus, each participant performed the experiment in a unique order. So, although order may have

impacted the results, we cannot separate out the effects of order from the effects of condition.

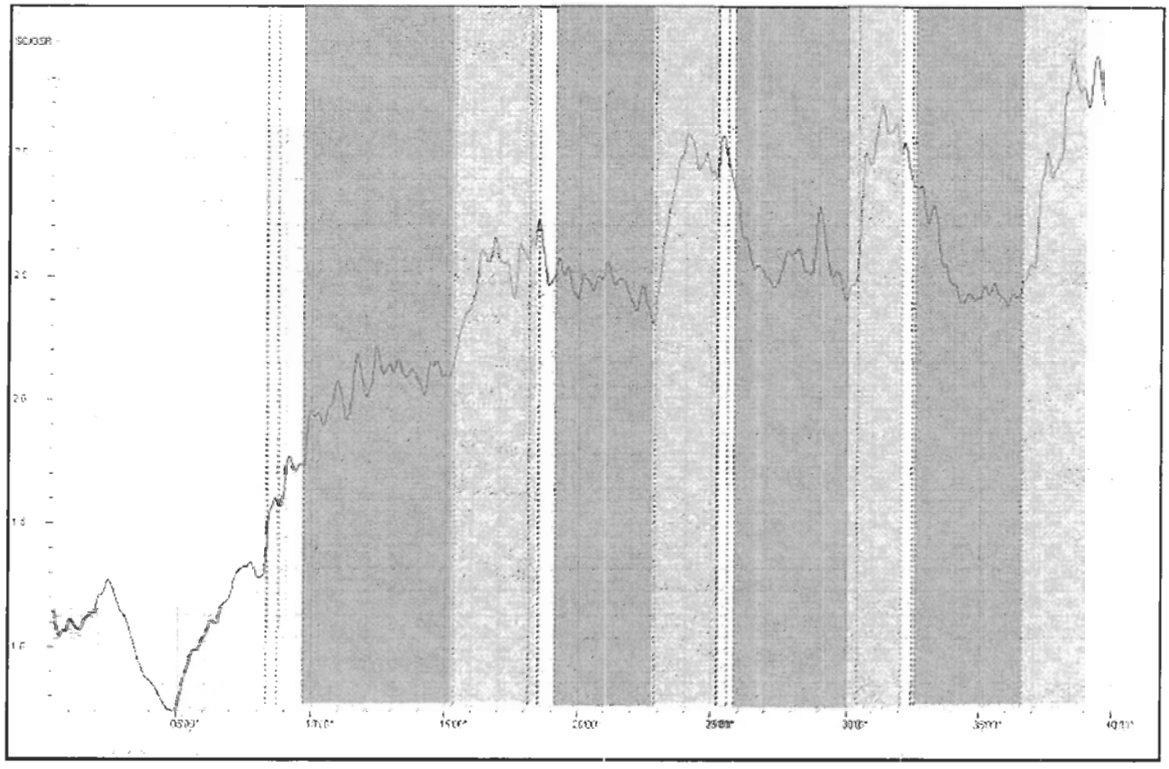


Figure 18: Participant 7's GSR signal over the course of the experiment. The areas shaded in light grey represent when the participant was being interviewed. The areas shaded in dark grey represent when the participant was playing the game.

6.7 Summary of Experiment One

Although we found many significant correlations for each individual, these correlations weren't consistent across participants. The main reason for the inconsistent results is likely the experimental manipulation that was chosen; however, there were also some methodological issues that contributed to irregular patterns of physiological activity. Primarily, the act of conducting the experiment produced

different phases in the experiment (e.g., play, interview, rest) that created greater physiological responses than the experimental manipulations themselves. In addition, the experimental manipulation that was chosen did not produce consistent subjective results across all participants. Without consistent subjective results, we cannot expect consistent physiological results. Given the data available, we cannot eliminate interview peaks, or change our experimental design to have a different control condition or a different experimental manipulation. Our sample size was also very small, but rather than add more participants to an imperfect experimental design, we took the methodological lessons learned and conducted a second experiment.

Chapter 7 EXPERIMENT TWO: TURING

We conducted a second study to further understand how body responses can be used to create an objective evaluation methodology. Because this methodology is a novel approach to evaluate play technologies, and the results from Experiment One were ambiguous, we used an experimental manipulation designed to maximize the difference in the experience for the participant. The participants played in two conditions: against another co-located player, and against the computer.

We chose these play conditions because we have previously observed pairs (and groups) of participants playing together under a variety of collaborative conditions [22, 54, 75, 120]. Our previous observations revealed that players seem to be more engaged with a game when another co-located player is involved. The chosen manipulation should yield consistent subjective results, and thus consistent physiological patterns of experience. Once we better understand how the body responds to play environments, more subtle manipulations could be explored.

Our goal was not to investigate whether there are differences between playing against a computer and a friend. We already know that the two play conditions are different. Our goal was to determine whether the physiological measurements could reveal differences between the two play conditions. As such, we called this experiment

Turing, since we were investigating whether we could use physiology to differentiate whether people were playing against a computer or a friend.

Our main suppositions for Experiment Two were that participants would be more excited, and would prefer playing against a friend over playing against a computer. Also, they would have more fun, and would be more engrossed in play against a friend. This preference would be reflected in their subjective experience as well as their physiological experience. Our previous studies on collaborative play, as well as the literature on physiology and emotion were used to generate the following experimental hypotheses.

H4: *Participants will prefer playing against a friend to playing against a computer. They will also find playing against a friend more fun, and engaging, and less boring.*

H5: *Participants will experience higher GSR values when playing against a friend than against a computer, a reflection of being more engaged, and having more fun.*

H6: *Participants will experience higher* EMG_{jaw} *values along the jaw when playing against a friend than against a computer, as a result of trying harder due to greater engagement.*

H7: *The differences in the participants' GSR signal in the two conditions will correlate to the differences in their subjective responses of engagement, fun, and/or excitement.*

Ratification of these hypotheses would provide support for two of our three main conjectures:

Conjecture A: *Physiological measures can be used to objectively measure a player's experience with entertainment technology.*

Conjecture B: *Normalized physiological measures of experience with entertainment technology will correspond to subjective reports.*

7.1 Participants

Ten male participants age 19 to 23 took part in the experiment. Participants were recruited from computer science and engineering students and recent graduates and were given a monetary remuneration of $20 for their participation. Before the experimental session, all participants filled out a background questionnaire (see Appendix 5). The questionnaire was used to gather information on their computer use, experience with computer and video games, game preference, console exposure, and personal statistics such as age and handedness.

All participants were frequent computer users. When asked to rate how often they used computers, nine subjects used them every day, and one subject used them often. The participants were also all self-declared gamers. When asked how often they played computer games, two played every day, seven played often, and one played rarely. The one participant who played computer games rarely, played console games occasionally. For the frequencies of responses to questions on computer usage and

game play, see Table 7 and Table 8. When asked how much they liked different game genres, role-playing was the favorite, followed by strategy games (see Table 9).

Table 7: Frequency of computer usage and game play from Experiment Two. Participants were asked to respond to how often they do each of the following activities:

	Never	*Rarely*	*Occasionally*	*Often*	*Every day*
Use computers?				1	9
Play computer games?		1		7	2
Play video (console) games?		2	3	4	1
Play computer/video games over the internet or network?		1	4	3	2
Play computer/video games with another co-located player?		3	3	3	1

Table 8: Frequency of computer usage and game play from Experiment Two. Participants were asked to respond to how much time they spend doing each of the following activities:

	Never	*< 3 hours a week*	*3-7 hours a week*	*1-2 hours a day*	*> 2 hours a day*
Use computers?				1	9
Play computer games?		3	3	1	3
Play video (console) games?	1	5	3		1
Play computer/video games over the internet or network?		4	4		2
Play computer/video games with another co-located player?	1	6	2		1

Table 9: Results of game genre preference from background questionnaires from Experiment Two. A 5-point Likert scale was used with "1" corresponding to "Dislike a lot" and "5" corresponding to "Like a lot".

	Mean	*St.Dev.*
Action	4.30	.68
Adventure	4.40	.84
Puzzle	3.50	1.1
Racing	3.80	.63
Roleplaying	4.90	.32
Shooting	4.10	.99
Simulation	4.30	.68
Sports	3.90	1.3
Strategy	4.78	.44

7.2 Play Conditions

Participants played NHL 2003 by EA Sports in both conditions (see Figure 12 for a screen shot). Two of the pairs were very experienced with the game, while the other three pairs were somewhat familiar or inexperienced with the game.

Participants played the game in two conditions: against another player, and against the computer. Participants were recruited in pairs so that they would be playing against friends rather than against strangers. Because they were recruited in pairs, one player competed against the computer before playing against their partner, while the other player competed against the computer after playing against their partner. This was to acknowledge effects due to the order of the presentation of conditions. Pairs who were somewhat inexperienced with NHL 2003 were given time to practice before the experiment in order to learn the controls.

Each play condition consisted of one 5-minute period of hockey. The game settings were kept consistent within each pair during the course of the experiment. All players used the Dallas Stars and the Philadelphia Flyers as the competing teams, as these two teams were comparable in the 2003 version of the game. All players used the overhead camera angle, and the home and away teams were kept consistent. This was to ensure that any differences observed within subjects could be attributed to the change in play setting, and not to the change in game settings, camera angle, or direction of play. The only difference between pairs was that experienced pairs played both conditions in a higher difficulty setting than non-experienced players.

7.3 Experimental Setting and Protocol

The experiment was conducted in a laboratory at Simon Fraser University. NHL 2003 was played on a Sony PS2, and viewed on a 36" television. A camera captured both of the players, their facial expressions and their use of the controllers, while an omnidirectional microphone captured the participants' comments. The game output, the camera recording, and the screen containing the physiological data were synchronized into a single quadrant video display, recorded onto tape, and digitized (see Figure 19). The quadrant video also contained the audio of the participants' comments, and the audio generated from the game. A diagram of the complete experimental setup can be seen in Figure 20.

Physiological data were gathered using the ProComp Infiniti system and sensors (see Figure 14), and BioGraph Software from Thought Technologies. Based on previous literature, we chose to collect galvanic skin response (GSR), electrocardiography

(EKG), electromyography of the jaw (EMG_{jaw}), and respiration. Heart rate (HR) was computed from the EKG signal, while respiration amplitude (RespAmp) and respiration rate (RespRate) were computed from the raw respiration data. We did not collect blood volume pulse data (BVP) because the sensing technology used on the finger is extremely sensitive to movement. As our subjects were operating a game controller, it wasn't possible to constrain their movements.

Upon arriving, participants signed a consent form (see Appendix 4). They were then fitted with the physiological sensors. One participant rested for 5 minutes, and then played the game against the computer. Both participants then rested for 5 minutes after which they played the game against each other. The second participant then rested again and played the game against the computer. When one participant was playing against the computer, the other participant waited outside of the room during the pre-play rest and the play condition. Because the participants were required to rest in the same room before playing each other, they wore headphones and listened to a CD containing nature sounds. This helped them to relax and ignore the other player in the room. They also listened to the CD when resting alone to maintain consistency. The resting period was included to allow the physiological measures to return to baseline levels prior to each condition. Experiment One showed that the act of filling out the questionnaires and communicating with the experimenter altered the physiological signals. The resting periods corrected for these effects. In order to utilize the resting periods as baseline controls, we would need much longer rest periods, and ensure that the nature sounds were indeed restful. We wanted to create an

environment that was as natural as possible, and extended periods of rest in between play conditions did not fit with this approach.

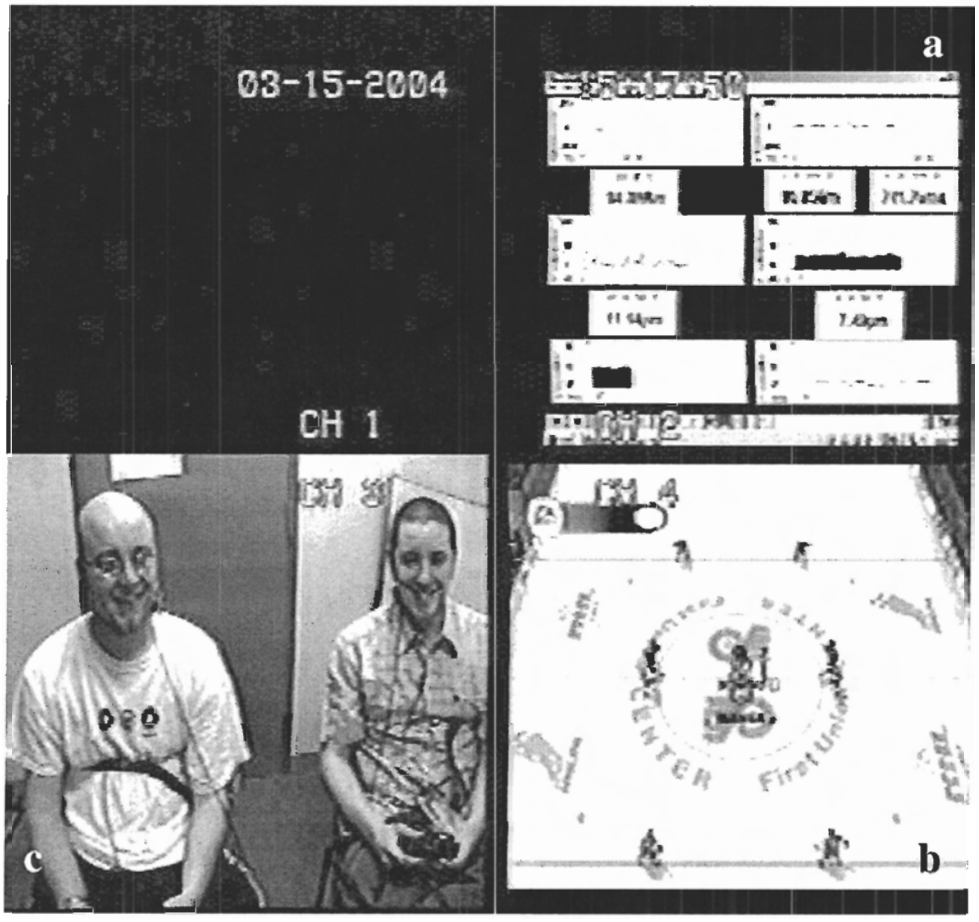


Figure 19: **Quadrant display for Experiment Two including: a) the screen capture of the biometrics, b) a screen capture of the game, and c) the camera feed of the participants. Audio of the participants' comments and audio from the game were included in the quadrant video.**

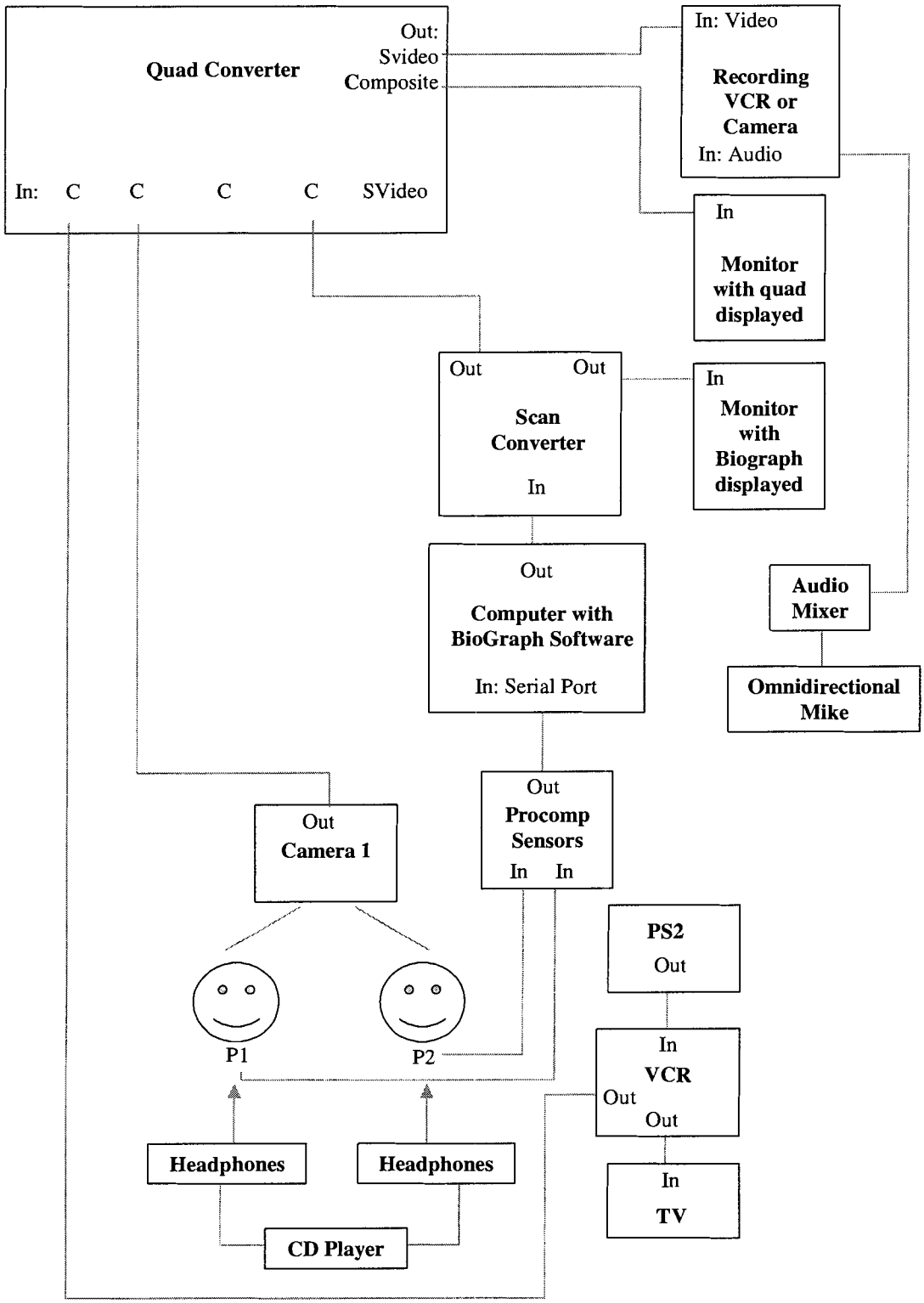


Figure 20: A diagram of the complete experimental set-up for Experiment Two.

After each condition, the participants filled out a condition questionnaire. The condition questionnaire contained their participant ID, the condition name, the level of play, and the final score (see Appendix 8). We also had subjects rate the condition using a Likert Scale. They were asked to consider the statement, "This condition was boring", rating their agreement on a 5-point scale with 1 corresponding to "Strongly Disagree" and 5 corresponding to "Strongly Agree". The same technique was used to rate how challenging, easy, engaging, exciting, frustrating, and fun that particular condition was. The questionnaire was filled out online using a laptop computer. Experiment One revealed that the physiological measurements for all participants reacted strongly to the interview process between each condition. We don't know what caused this effect but feel that the act of speaking and answering questions may have contributed. As a result, we chose to have participants fill out html-based questionnaires using a laptop computer, and then rest again for 5 minutes. After completing the experiment, subjects completed a post-experiment questionnaire (see Appendix 9). We asked them to decide in retrospect which condition was more enjoyable, more fun, more exciting, and more challenging. They were also asked in which condition they would choose to play, given the choice to play against a friend or against the computer. Discussion of their answers was encouraged. The experimenter verbally administered the post-experiment questionnaire.

7.4 Data Analyses

The subjective data from both the condition questionnaires and the post experiment questionnaires were collected into a database, and analyzed using non-parametric statistical techniques.

In terms of the physiological data, EKG data were collected at 256 Hz, while GSR, respiration, and EMG_{jaw} were collected at 32 Hz. HR, RespRate, and RespAmp were computed at 4 Hz. Physiological data for each rest period and each condition were exported into a file. As in Experiment One, noisy EKG data may produce heart rate (HR) data where two beats have been counted in a sampling interval or one beat has only been counted in two sampling intervals. We inspected the HR data and corrected these erroneous samples. For each condition and rest period, HR data were then computed into the following measures: mean HR, peak HR, min HR, and standard deviation of HR. The same four measures (mean, peak, min, and standard deviation) were also computed on the GSR data, EMG_{jaw} data, RespAmp data, and RespRate data.

7.5 Results and Discussion

Results of the subjective data analyses are described first, followed by results of the physiological data analyses. Finally, correlations between the subjective data and the physiological data are presented.

7.5.1 Subjective Responses

In Experiment One, our experimental setting seemed to have impacted the results more than our experimental manipulations. Although we addressed these issues, to be certain of our results, we wanted to closely examine any potential methodological problem. We used the chi-squared statistic to determine whether subjective responses were influenced by order of presentation of condition or outcome of the condition (win, loss, or tie). There were no significant effects of order on any of the subjective

measures, either on the condition questionnaire, or on the post-experiment questionnaire. There was a significant effect of condition outcome on boredom rating, when participants played against the computer. Participants who lost to the computer rated the condition as significantly more boring (mean=4.0, N=2) than subjects who beat the computer (mean=2.0, N=5), or who tied the computer (mean=1.67, N=3) ($\chi 2$=12.38, p=.015). However, there was no difference in boredom ratings depending on game outcome when participants played against a friend (mean(win)=1.67, N=3, mean(loss)=2.0, N=3, mean(tie) =1.5, N=4) ($\chi 2$=4.50, p=.343, see Figure 21). The game outcome had no significant impact on any of the other subjective measures.

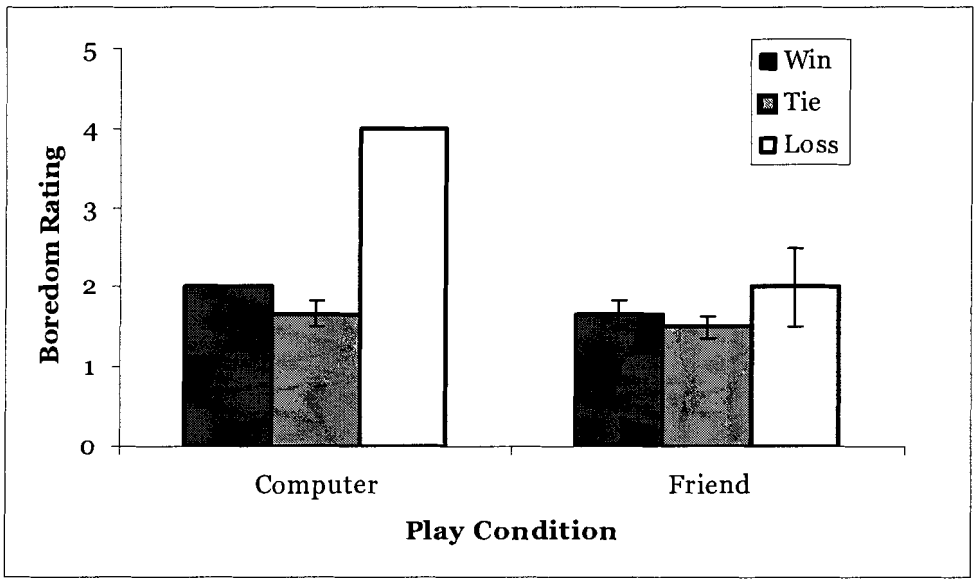


Figure 21: **Mean subjective ratings (± SE) for boredom in Experiment Two, separated by game outcome (win, loss, tie). Participants who lost to the computer rated the condition as significantly more boring than those who beat or tied the computer. Boredom rating when playing against a friend was not impacted by game outcome. Note that the standard error for those who beat (N=5) or lost (N=2) to the computer was zero.**

In addition, the ratings for playing against the computer were compared to the ratings for playing against a friend. Players found it significantly more boring (χ^2=4.0, p=.046) to play against a computer than against a friend, and significantly more engaging (χ^2=4.0, p=.046), exciting (χ^2=6.0, p=.014), and fun (χ^2=6.0, p=.014) to play against a friend than a computer (Friedman test for two related samples). See Figure 22, and Table 10 for a synopsis of these results.

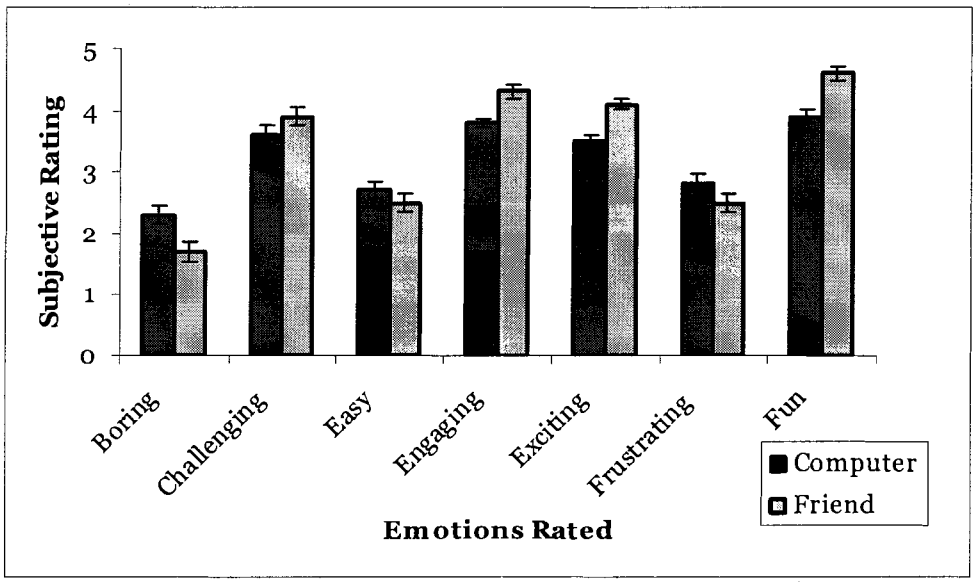


Figure 22: Mean subjective ratings (± SE) for Experiment Two, separated by condition. Subjects were asked to rate each experience state on a 5-point scale. Identifying strongly with that experience state is reflected in a higher mean. Participants found it significantly more boring to play against the computer, and significantly more engaging, exciting, and fun to play against a friend.

On the post-experiment questionnaire, when asked whether it was more enjoyable to play against the computer or a friend, all 10 subjects chose playing against a friend. All 10 subjects also stated that it was more fun and more exciting to play against a

friend; however, half of the subjects thought it was more challenging to play against the computer. When those five participants were asked why it was more challenging to play against the computer, most felt that their partner was not as good of a player as the computer. The five participants who were more challenged by their partner felt that the computer was too predictable. The participants were grouped into two "challenge groups", depending on whether they felt more challenged by their friend or the computer. When asked if given a choice, in which condition they would choose to play, all 10 subjects reported that they would choose to play against a friend.

It isn't surprising that the participants found the game fun, and that they enjoyed playing against a friend more than the computer. When recruiting players, we asked that they be computer game players familiar with a game controller, drawing people that generally enjoy playing computer games (as seen in the results from the background questionnaire). We recruited the participants individually, but asked them to bring their own partner. We didn't want the participants playing against strangers, which may have discouraged people who prefer playing alone from signing up.

Our first experimental hypothesis stated that participants would prefer playing against a friend to playing against a computer. The described subjective results confirm this hypothesis.

Table 10: Results of condition questionnaires for Experiment Two. Subjects were asked to rate each experience state on a 5-point scale. Identifying strongly with that experience state is reflected in a higher mean.

	Playing against computer		Playing against friend		Difference between conditions	
	Mean	*St.Dev.*	*Mean*	*St.Dev.*	χ^2	*p*
Boring	**2.3**	**.949**	**1.7**	**.949**	**4.0**	**.046**
Challenging	3.6	1.08	3.9	.994	1.8	.180
Easy	2.7	.823	2.5	.850	1.0	.317
Engaging	**3.8**	**.422**	**4.3**	**.675**	**4.0**	**.046**
Exciting	**3.5**	**.527**	**4.1**	**.568**	**6.0**	**.014**
Frustrating	2.8	1.14	2.5	.850	.67	.414
Fun	**3.9**	**.738**	**4.6**	**.699**	**6.0**	**.014**

7.5.2 Physiological Measures

Each physiological measure was computed into means for each participant. Means for the physiological data (GSR, EMG_{jaw}, HR, RespRate, and RespAmp) were analysed using a repeated measures multivariate analysis of variance (MANOVA) with play conditions as the independent variable, and the five physiological signals as dependent variables. Order of presentation and challenge group (as identified in the post-experiment questionnaire) were included as factors to determine whether there were effects due to order of condition, and to differentially analyze the physiological results for the two different challenge groups identified in the post-experiment questionnaire. There were no significant main effects of order, or any interactions between the play condition and the order in which it was presented. Thus, the resting period between play conditions served the purpose of returning the physiological measures to a baseline state. We also examined whether game outcome (win, loss, tie) differentially

affected the participants' physiological measures. There were no systematic effects of game outcome on any of the physiological measures analysed.

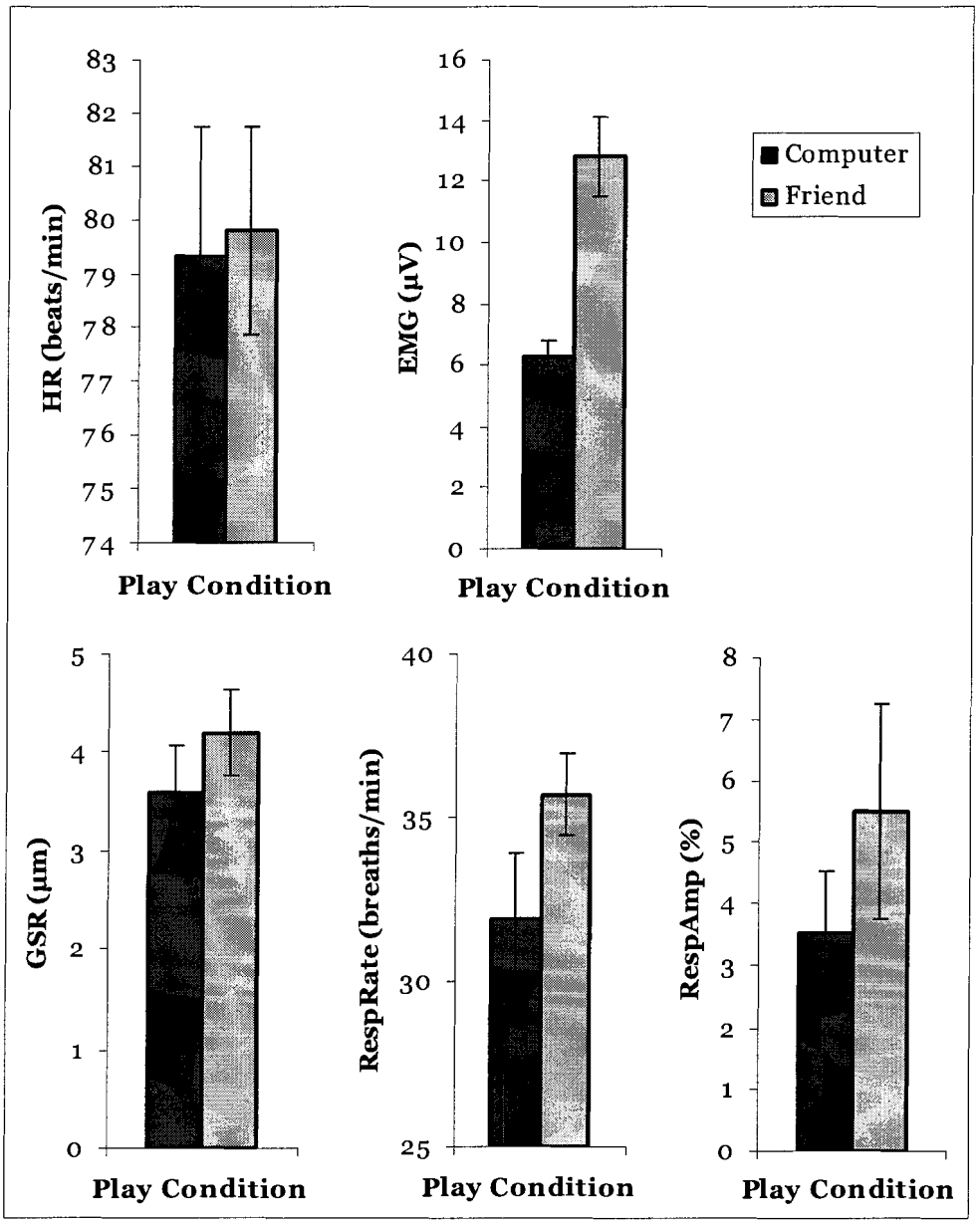


Figure 23: **Mean physiological results (± SE) separated by play condition. GSR and EMG_{jaw} were significantly higher when playing against a friend. There was no difference in HR, RespAmp, or RespRate between conditions. Note that the error bars are exagerrated since there are large individual differences in physiological measures, and these values are averaged over the group.**

Our second experimental hypothesis assumed that GSR would be greater when playing against a friend as compared to playing against the computer, due to greater engagement. Figure 23 shows how mean GSR was significantly higher when playing against a friend (mean=4.19μm, SD =3.0) as compared to playing against a computer (mean=3.58 μm, SD=2.8), ($F_{1,5}$ =7.4, p=.042, η^2=.60). Because of the individual variability in physiological data, the standard deviations are quite high; however, the average increase in GSR when playing against a friend was 36% of the signal span (using the resting value of GSR as the lower bound and the maximum GSR value during the experiment as the upper bound). Also, the partial eta-squared value of .60 reveals that 60% of the total variability in the measure can be attributed to play condition.

In addition, when examined individually, Figure 24 shows how the pattern of higher GSR when playing a friend was consistent for 9 of the 10 subjects, which is a significant trend (Z=2.4, p=.017). The one participant whose GSR did not increase was also the only participant who did not increase his subjective rating for fun when playing against a friend, and as such, we would not expect his GSR to be higher when playing against his friend. He felt more challenged playing against the computer than against his partner (challenge(computer) = 5, challenge(friend) = 2). He also felt that it was easier to play against his partner than the computer (easy(computer) = 2, easy(friend) = 4)). Throughout the experiment, his partner had difficulty learning the controls to the game. This circumstance could have created an anomalous play experience against his friend, and explain his lower GSR.

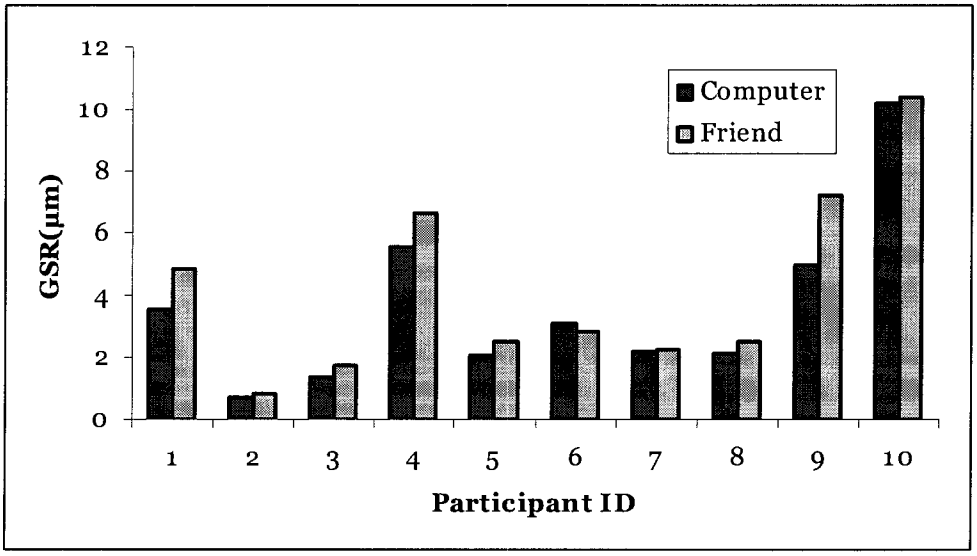


Figure 24: Mean GSR values (µm) for Experiment Two, separated by participant and play condition. Mean GSR was higher when playing against a friend as compared to playing against a computer. This pattern was seen in all players with the exception of participant 6.

Our third hypothesis states that we expected EMG activity along the jaw to be greater when playing against a friend, as we expected participants to try harder and be more competitive when playing against a friend, due to greater engagement. Although we placed the surface EMG on the jaw to collect data on tension in the jaw, these results are likely overshadowed by interference from smiling and laughing. We cannot separate out these effects, to determine the EMG scores for jaw clenching alone. With this in mind, mean EMG_{jaw} was significantly higher when playing against a friend (mean=12.8 µV, SD=8.2) as compared to playing against a computer (mean=6.3 µV, SD=3.3), ($F_{1,5}$ =14.8, p=.012, η^2=.75, see Figure 23). The factor of condition accounts for 75% of the variability in the measure, and Figure 25 shows how the increase was consistent for 9 of the 10 subjects, which is a significant trend (Z=2.7, p=.007).

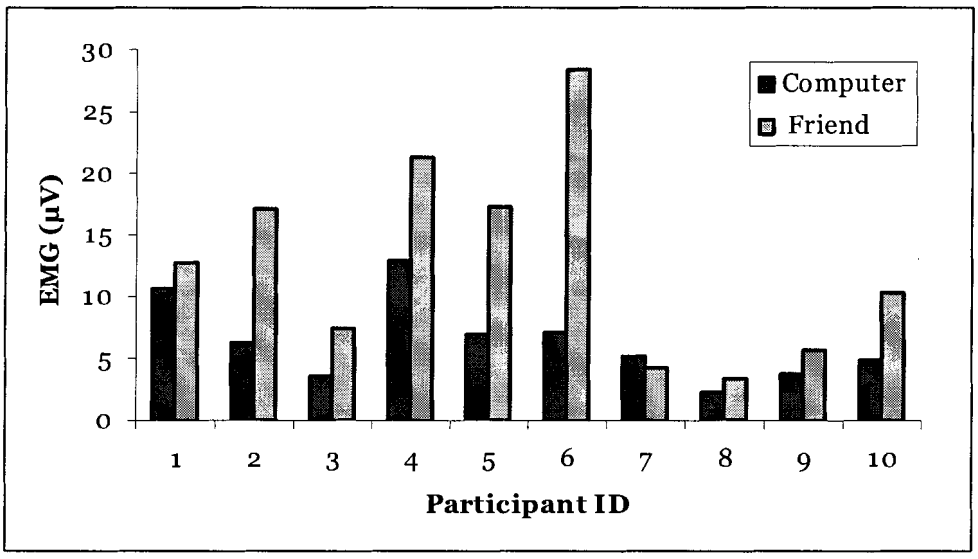


Figure 25: Mean EMG_{jaw} values for Experiment Two, separated by participant and play condition. Mean EMG_{jaw} was higher when playing against a friend as compared to playing against a computer. This pattern was seen in all players with the exception of participant 7.

Based on psychophysiological theories, we didn't expect to see any differences between the conditions in heart rate (HR), respiratory amplitude (RespAmp), or respiratory rate (RespRate). The MANOVA showed no significant differences in HR, RespAmp, or RespRate between the two play conditions (HR: $F_{1,5}$ =1.58, p=.264, η^2=.24; RespAmp: $F_{1,5}$ =2.15, p=.202, η^2=.30; RespRate $F_{1,5}$ =.69, p=.444, η^2=.121, see Figure 23).

In the post-experiment questionnaires, half of our participants felt that playing against the computer was more challenging, and half felt that playing against their friend was more challenging. As such, we included this grouping as a between subjects factor in our MANOVA on the physiological data to investigate whether the perception of challenge differentially affected the physiological measures, as shown in Figure 26.

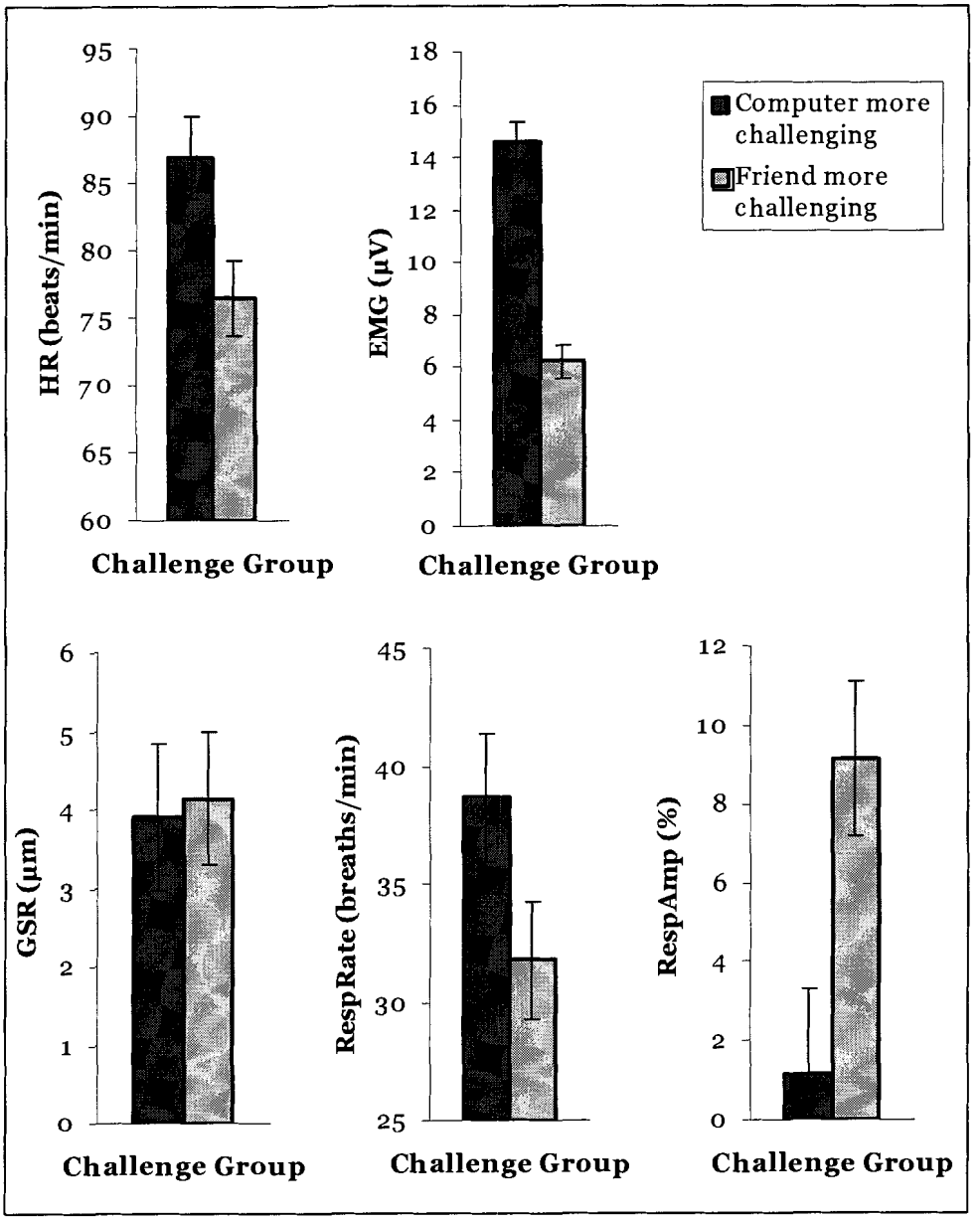


Figure 26: **Mean physiological results (± SE) separated by challenge group. EMG_{jaw} was significantly higher for the group that felt more challenged when playing against a computer. This effect did not interact with play condition. There was no difference in GSR, HR, RespAmp, or RespRate between challenge groups. Note that the error bars are exagerrated since there are large individual differences in physiological measures, and these values are averaged over the group.**

There was a main effect of challenge group on EMG_{jaw}. Those who felt that playing against the computer was more challenging had a higher mean EMG_{jaw} over both play

conditions (mean=14.6 μV, SE=1.4) than those who felt that playing their friend was more challenging (mean=6.2 μV, SE=1.3) ($F_{1,5}$ =19.4, p=.007, η^2=.80). This effect did not interact with play condition. The MANOVA showed no significant differences in GSR, HR, RespAmp, or RespRate between the two challenge groups (GSR: $F_{1,5}$=0.009, p=.928, η^2=.002; HR: $F_{1,5}$=1.55, p=.268, η^2=.24; RespAmp: $F_{1,5}$=1.87, p=.229, η^2=.27; RespRate $F_{1,5}$=.86, p=.397, η^2=.15).

7.5.3 Physiological Measures as a Continuous Data Source

The comparison between the means for two conditions provides a good basis for using physiological measures as an objective indicator of experience with entertainment technology. However, we can't say with any degree of certainty whether the tonic level is raised, or whether there are more phasic responses[7]. As such, in addition to comparing the means from the two conditions, we investigated GSR responses for individual events. One of the advantages of using physiological data to create evaluation metrics is that physiological signals provide high-resolution, continuous, contextual data. GSR is a highly responsive body signal, it provides a fast-response time-series, reactive to events in the game. To inspect GSR response to specific events, we chose to examine small windows of time surrounding goals scored and fights in the game. Goal events were windowed for 10 seconds before scoring and 15 seconds after scoring, in order to establish a pre-event level as well as contain an entire potential GSR response to a goal. There were 10 instances where participants

[7] Tonic activity refers to the baseline measure of a system; the background or resting level of the activity of a particular physiological measure. Phasic activity refers to a discrete response to a stimulus, or an evoked response. Phasic activity can be either an increase or a decrease in frequency, amplitude, or latency [123].

scored in both play conditions. All of these participants experienced a significantly larger GSR response to goals scored against another player versus goals scored against the computer (t_4=6.7, p=.003). The magnitude of the response was calculated as the span of the response (peak minus min) during the windowed time period. An example of one participant's result scoring against the computer twice and against a friend once is shown in Figure 27.

When two players begin a hockey fight, the game cuts to a different scene and the players throw punches using buttons on the controller (see Figure 28). Fight sequences were analysed from the time when the pre-fight cut scene began to when the post-fight cut scene ended. There were three instances of participants who participated in hockey fights both against the computer and against their friend. One participant won both fights, one lost both, and one won against the computer and lost against their friend. Even so, all participants exhibited a significantly larger response to the fight against the friend than the fight against the computer (t_2=6.0, p=.027). An example of one player's response to a fight sequence against the computer and against a friend is shown in Figure 29.

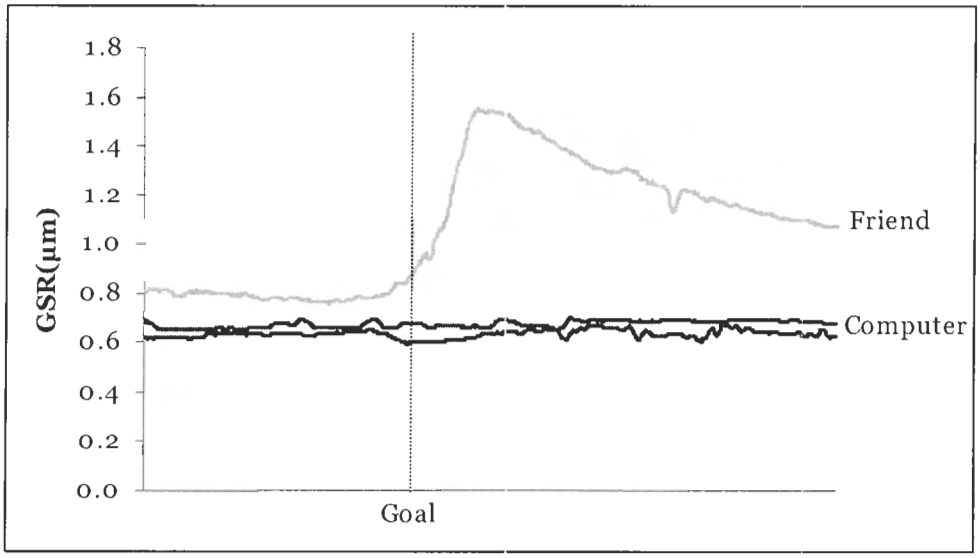


Figure 27: Participant 2's GSR response to scoring a goal against a friend and against the computer twice. Note the much larger response when scoring against a friend. Data were windowed 10 seconds prior to the goals and 15 seconds after.

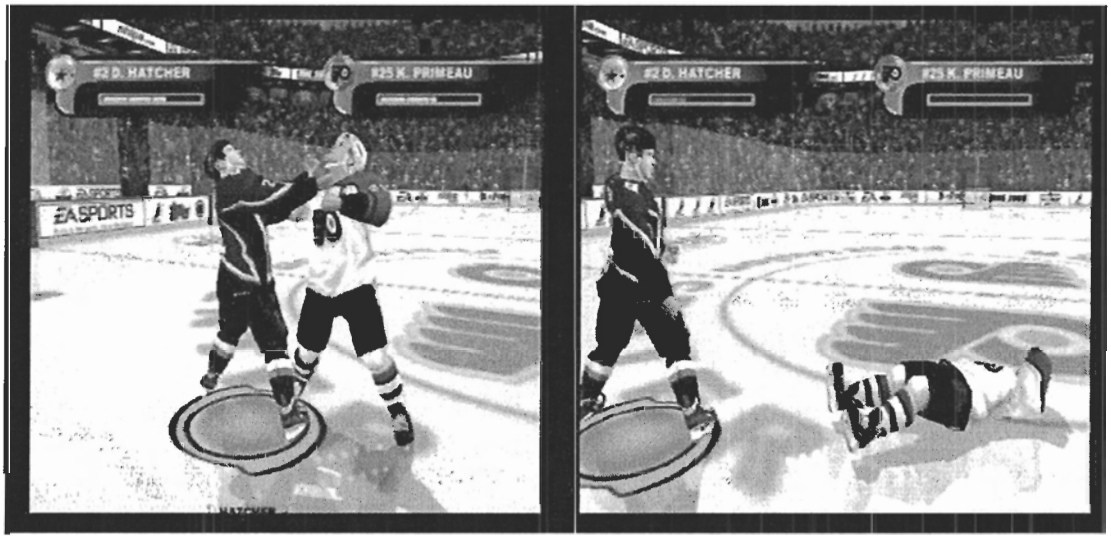


Figure 28: Fight sequence in NHL 2003 by EA Sports. The first frame shows the players in a fight. The second frame is after the Dallas Stars player won.

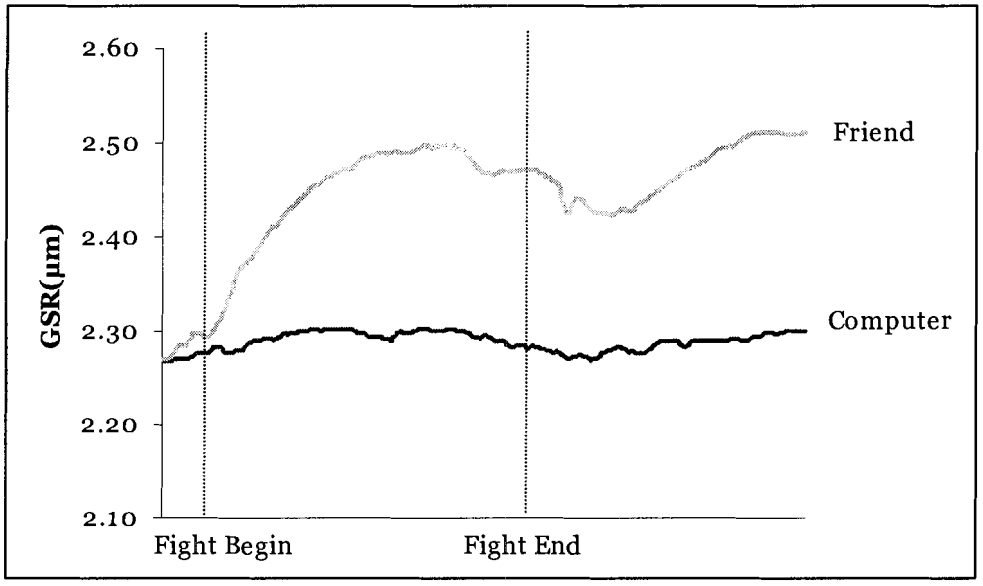


Figure 29: Participant 9's GSR response to engaging in a hockey fight while playing against a friend versus playing against the computer.

7.5.4 Correlation of Subjective Responses and Physiological Data

Since physiological data has very large individual differences, and individual baselines have to be taken into account, we could not directly compare the means of the time-series data to the results from the subjective data from the condition questionnaires. In previous literature, (see [74] for an overview), researchers have rarely correlated physiological data to other types of data. One exception is Vicente et al. [137] who normalized HRV and compared subjective ratings to normalized HRV.

In Experiment One, we correlated physiological results to subjective results for each individual and then determined whether these patterns were consistent across individuals. In this case, we only have two conditions (friend and computer),

rendering this method unusable, since with only two conditions, correlations will either be zero or one depending on the direction of the differences.

In order to perform a group analysis, we transformed both the physiological and subjective results into dimensionless numbers between zero and one. For each player, the difference between the conditions was divided by the span of that individual's results. The physiological data were converted using the following formula:

$$\text{Physiological}_{\text{Normalized}} = \frac{\text{Mean}C - \text{Mean}F}{\text{MAX}\{\text{Peak}C\text{-Min}C, \text{Peak}F\text{-Min}F\}}$$

where *C* refers to playing against the computer and *F* refers to playing against a friend.

The subjective results were handled similarly:

$$\text{Subjective}_{\text{Normalized}} = \frac{C - F}{4}$$

These normalized measures were then correlated across all individuals. We weren't interested in how the subjective results correlated with each other. For example, it is to be expected that boredom will be negatively related to excitement. Similarly, we didn't correlate physiological measures with other physiological measures. All correlations between subjective measures and physiological measures are shown in Table 11.

Table 11: Correlations between normalized subjective measures and normalized physiological measures. Significant correlations (p, 2-tailed) are shaded in grey. r values are Pearson correlation coefficients.

		GSR	HR	RespAmp	RespRate	EMG_{jaw}
Fun	r	.694	-.156	-.173	.086	-.608
	p	.026	.667	.633	.812	.062
Boredom	r	-.508	-.218	.249	-.182	.823
	p	.134	.545	.487	.614	.003
Challenge	r	-.383	-.173	.697	.069	.782
	p	.275	.632	.025	.850	.008
Ease	r	.489	-.068	-.684	-.534	-.649
	p	.151	.852	.029	.112	.042
Engagement	r	.160	.248	.259	-.006	.056
	p	.659	.489	.470	.988	.878
Excitement	r	.469	-.272	.381	.333	-.124
	p	.172	.447	.277	.348	.733
Frustration	r	-.641	.259	.041	-.638	.181
	p	.046	.470	.910	.047	.617

Since mean GSR was higher when playing against a friend, and participants also rated this condition as more fun and exciting, we hypothesized that a correlation between GSR and fun, excitement, or boredom might exist. By themselves, the subjective and physiological results reveal that a participant's GSR is higher in a condition that they also rate as more fun. A correlation of the normalized differences would show that the *amount* by which subjects increased their fun rating when playing against a friend is proportional to the *amount* that GSR increased in that condition. Using Pearson's coefficient, we found that normalized GSR was correlated with normalized fun (r=.69,

p=.026). Thus, the level of arousal experienced by the subjects corresponded with their subjective reported experience of fun (see Figure 30). We also found that normalized GSR was inversely correlated with normalized frustration (r=.64, p=.046). Thus, the amount by which their GSR decreased when playing against the computer is comparable to the increased amount in their frustration rating.

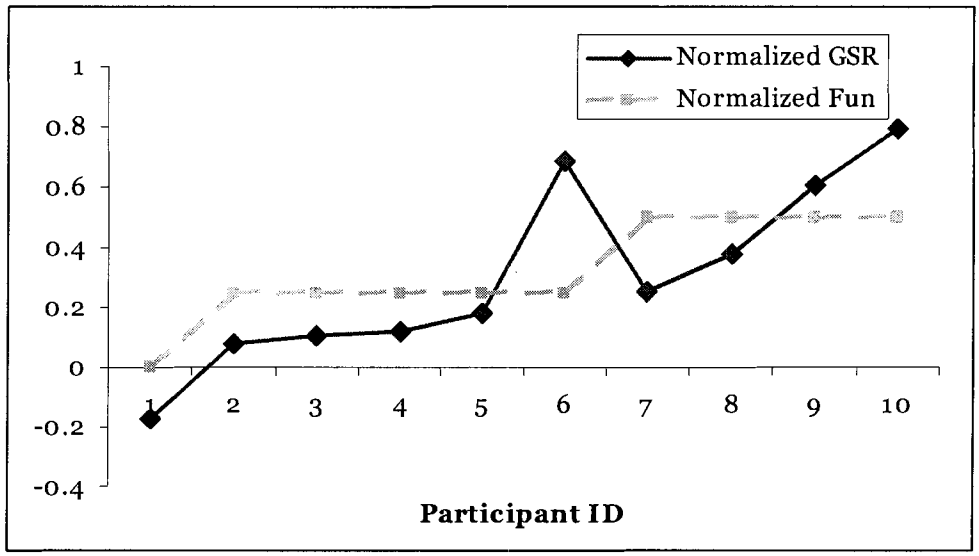


Figure 30: Normalized GSR is correlated with normalized fun (r = .70, p =.026).

We also found that normalized respiratory amplitude was correlated with normalized challenge (r=.70, p=.025) and inversely correlated with normalized ease (r=.68, p=.029). We had previously seen the Challenge-RespAmp correlation in Experiment One when observing people playing NHL 2003 in different difficulty levels. In the present experiment, respiration amplitude increased for all 10 participants when playing against a friend, although this result was non-significant. Although half the participants said in the post-experiment questionnaire that playing against the computer was more challenging, the condition questionnaires revealed that 9 of the 10

subjects rated the challenge of playing against a friend as the same or higher than playing against the computer.

Normalized respiration rate was inversely correlated with frustration (r=.64, p=.047). Respiration rate tends to increase with emotional arousal, so we might expect that an aroused state of frustration would increase respiration rate; however, the frustration that players were experiencing with the controls might have caused them to 'shut down' rather than become more aroused. In our experiment, participants were neither encouraged, nor discouraged to talk, but it seemed that there was more talking and laughing when playing against a friend than when playing against a computer. Given that talking and laughing affect respiration, results involving respiration need to be interpreted with caution.

Normalized EMG_{jaw} correlated with boredom and challenge, (r=.82, p=.003; r=.78, p=.008) and inversely with ease (r=.64, p=.042). We would expect the mean increase in jaw clenching to correspond to an increase in challenge and a decrease in ease since people clench their jaws when concentrating. The boredom correlation is a little surprising since we would expect a bored participant to be more relaxed; however, since boredom was indexed to game outcome when playing against the computer (section 7.5.1), those same participants could have been clenching their jaw in concentration trying to beat the computer. Although the EMG sensors were placed to sense jaw clenching, there may have been interference from smiling and laughing, so these results need to be interpreted with caution.

There were no significant correlations between heart rate and any of the subjective measures.

7.6 How Issues From Goldilocks were Addressed

Although our approaches to solving the methodological issues uncovered in Experiment One are described throughout Sections 7.1 through 7.4, they are reiterated in this section. The four issues from Goldilocks and how we altered the design and analysis of Experiment Two follow:

Subjects enjoyed playing in all conditions: We chose a new experimental manipulation. Previous studies revealed a different subjective experience when playing against a co-located friend than playing alone. Although this manipulation seemed obvious in terms of the research contribution, it allowed us to examine physiological responses in an experiment where we expected homogeneous subjective reports across participants.

Variability inherent in game play: We continued to collapse the time series physiological data into a single data point through averaging. In addition, we examined windows of data surrounding interesting game events. This formed the basis for our investigation into the use of physiological measures as a source of continuous data.

High resting baseline: The participants listened to a relaxation CD during the resting periods. This helped them to relax and not be as aware of the experimental surroundings. In addition, they rested for longer than in Experiment One. Participants

also rested prior to each condition allowing their physiological measures to return to baseline levels.

Interview effects: In Experiment One, the interview process had greater impact on their physiological measures than the experimental manipulations. As a result, we chose to administer the questionnaires online, without the presence of an interviewer. We also introduced a resting period prior to each condition in order to allow any artificially elevated signals to return to baseline levels.

Order effects: In Experiment One, the interview process raised the participants' GSR signals, which caused a steep upwards drift over the experiment. The rest periods in between conditions allowed the physiological measures to return to resting values. Also, the use of two conditions allowed us to evaluate order as a between subjects factor, and we determined that order did not impact any of the physiological measures or interact with play condition.

7.7 Summary of Experiment Two

After addressing our methodological issues from Experiment One, Experiment Two tested and supported four experimental hypotheses:

H4: *Participants preferred playing against a friend to playing against a computer. They also found playing against a friend more fun, and engaging, and less boring.*

H5: *Participants experienced higher GSR values when playing against a friend than against a computer, a reflection of being more engaged, and having more fun.*

H6: *Participants experienced higher EMG_{jaw} values along the jaw when playing against a friend than against a computer, possibly as a result of trying harder due to greater engagement.*

H7: *The differences in the participants' GSR signal in the two conditions correlated to the differences in their subjective responses for fun and/or excitement.*

We also found other correlations between the normalized subjective measures and the normalized physiological measures.

The ratification of these hypotheses, along with the other results, provide support for our first two conjectures:

Conjecture A: *Physiological measures can be used to objectively measure a player's experience with entertainment technology.*

Conjecture B: *Normalized physiological measures of experience with entertainment technology will correspond to subjective reports.*

Normalizing and correlating the data is a powerful tool because it shows that the *amount* by which participants increased their subjective ratings corresponded to the *amount* by which their mean physiological data increased. In addition, this approach contains results that may otherwise get lost. For example, we saw in section 7.5.2 that participant 6's GSR decreased when playing against a friend. Further inspection revealed that he was the only participant who didn't increase his rating of fun when playing against a friend. Figure 30 shows how this explanation is encompassed in the

normalization and correlation technique. The ANOVAs show results when all participants are responding in a similar manner, however the normalization and correlation will reveal patterns even when participants are responding differently from one another, a useful tool when investigating something as individual as engagement with play technologies.

The confirmation of our hypotheses provided support for our two main conjectures: that physiological measures can be used as objective indicators for the evaluation of co-located, collaborative play; and that the normalized physiological results will correspond to subjective reported experience.

Subjective data yield valuable quantitative and qualitative results. However, when used alone, they do not provide sufficient information. In game design, reward and pacing are important features. Utilizing a single subjective rating can wash out this variability, since subjective ratings provide researchers with a single data point representing an entire condition. Think-aloud techniques [90], which are popular for use in productivity systems cannot effectively be used with entertainment technology because of the disturbance to the player, and the impact they have on the condition itself. In pilot testing, we employed a retrospective think-aloud technique, conducted while playing back the condition to the participant. Although informative, this technique qualifies the experience, rather than providing concrete quantitative data. In addition, the think-aloud process does not occur within the context of the task, but in reflection of the task. Finally, we found that participants were very good at reported

what happened in the game, but were very bad at reporting what they felt about what happened.

This experiment showed that when physiological data are analysed into averages for each condition, they yield meaningful results that respond in a similar manner to subjective reports. These results have the same disadvantage as subjective results, in that they are single points of data representing an entire condition; however, unlike subjective reporting, they represent an *objective* measure of user experience. Used in concert, these two methods can provide a more detailed and accurate representation of the player's experience.

The raised GSR signals when playing against a friend reveal that players are more aroused when playing against a friend than when playing against a computer. However, we do not know whether this elevated result can be attributed to a higher tonic level or more phasic responses. Physiological data provides a high-resolution time series, responsive to player experience. Using methods like the time-window analysis presented here provides continuous objective data that can be used to evaluate the player experience, yielding salient information that can discriminate between experiences with greater resolution than averages alone. In this experiment, we graphically represented continuous responses to different game events, and looked at the magnitude of the response using the span of the physiological measure. In the next experiment, we propose to take advantage of the high-resolution, contextual nature of physiological data to provide an objective, continuous measure of player experience. Based on the results of Experiment Two, we believe that physiological

metrics can be used to model user emotional experience when playing a game; providing continuous and objective metrics of emotion.

Chapter 8 EXPERIMENT THREE: CONTINUOUS EVALUATION OF EMOTION STATE

We conducted a third study to investigate whether we could model emotional responses to play technologies, creating an objective and quantitative method of evaluation. Successful results would provide support for our last conjecture:

Conjecture C: *Physiological metrics can be used to model user emotional experience when playing a game, providing continuous, quantitative, and objective metrics of evaluation for interactive play technologies.*

Because of the success of the experimental manipulation used in Experiment Two, we continued to use the manipulation of the playing partner to create different experimental conditions. The participants played in three conditions: against a co-located friend, against a co-located stranger, and against the computer. We added the stranger condition to yield more information on how play condition affects the gaming experience. As with our previous experiments, we were not interested in whether there was a difference between playing against a friend, a stranger, or a computer. We have observed many groups of people playing with interactive technologies, and we know that these three play conditions yield very different play experiences; rather, we were

interested in whether our model of emotion could detect the differences between the conditions.

Modeling emotions could be a powerful evaluative tool because modeled emotions are quantitative and objective, filling the knowledge gap for evaluating entertainment technologies identified in section 2.3.4. In addition, modeled emotions could be represented continuously over a session, drastically increasing the evaluative bandwidth over current techniques.

We used normalized GSR, HR, $EMG_{smiling}$, and $EMG_{frowning}$ signals as inputs to a fuzzy logic model. To generate values for user emotion, we modeled the data in two parts. First, we computed arousal and valence values from the normalized physiological signals, and then we used these arousal and valence values to generate emotion values for boredom, challenge, excitement, frustration, and fun.

8.1 Experimental Details

The details in the section apply to data that was collected for 12 participants. Six of the participants were used to generate the emotion models, which are described in this chapter. The remaining six participants were used to validate the modeled emotions by comparing the results to reported emotions through subjective responses. The validation is discussed in Chapter 9.

8.1.1 Participants

Twenty-four male participants aged 18 to 27 took part in the experiment. Participants were recruited from university undergraduate and graduate students. Participants were

recruited in pairs to ensure that they would be playing against a stranger in only one of the co-located conditions. We wanted all of the participants to be independent subjects, statistically unrelated to any of the other participants, so we only treated one player in each pair as the participant. As such, we designed the experiment for 12 participants in 12 pairs, and we report data for 12 participants; one member of each pair.

Before the experimental session, all participants filled out a background questionnaire (see Appendix 5). The questionnaire was used to gather information on their computer use, experience with computer and video games, game preference, console exposure, and personal statistics such as age and handedness.

All participants were frequent computer users. When asked to rate how often they used computers, all 12 subjects used them every day. When asked how often they played computer games, one played every day, four played often, three played occasionally, and four played rarely. When asked how often they played video (console) games, two played every day, three played often, four played occasionally, two played rarely, and one never played console games. The one participant who never played video games replied that he occasionally played console games. For the frequencies of responses to questions on computer usage and game play, see Table 12 and Table 13. When asked how much they liked different game genres, action was the favorite, followed by sports, and adventure games (see Table 14).

Table 12: Frequency of computer usage and game play. Participants were asked to respond to how often they do each of the following activities. Note that it was two different participants who replied never to playing computer and video games.

	Never	*Rarely*	*Occasionally*	*Often*	*Every day*
Use computers?					12
Play computer games?	1	2	4	2	3
Play video (console) games?	1	2	3	5	1
Play computer/video games over the internet or network?	2	2	3	3	2
Play computer/video games with another co-located player?	2		6	3	1

Table 13: Frequency of computer usage and game play. Participants were asked to respond to how much time they spend doing each of the following activities. Note that it was two different participants who replied never to playing computer and video games.

	Never	*< 3 hours a week*	*3-7 hours a week*	*1-2 hours a day*	*> 2 hours a day*
Use computers?*				2	9
Play computer games?	1	3	2	4	2
Play video (console) games?	1	5	3	2	1
Play computer/video games over the internet or network?	3	5		3	1
Play computer/video games with another co-located player?	2	7	2	1	

* missing one data point

Table 14: Results of game genre preference from background questionnaires. A five-point Likert scale was used with "1" corresponding to "Dislike a lot" and "5" corresponding to "Like a lot".

	Mean	*St.Dev.*
Action	4.2	0.7
Adventure	4.1	0.7
Puzzle	3.8	1.1
Racing	3.7	1.2
Roleplaying	3.9	1.5
Shooting	3.8	0.9
Simulation	3.3	1.4
Sports	4.1	1.2
Strategy	3.8	1.5

8.1.2 Play Conditions

Participants played the game in three conditions: against a co-located friend, against a co-located stranger, and against the computer. Order of the presentation of the conditions was fully counterbalanced. The stranger remained constant for all participants, and was a 29 year-old male gamer, who was instructed to match each participant's level of play to the best of his ability.

Because we recruited participants in pairs, and were only treating one member of the pair as the participant, we needed to decide at the beginning of the session which player we would test. When the participants arrived, we chose the person with the least amount of facial hair to be the participant. If we couldn't discriminate between participants using facial hair, we took the player who was wearing a hat. If this didn't discriminate, we took the player who entered the room first.

For pairs that were tested in the friend condition first, we began with both players in the room, asking the non-participant to wait outside of the experiment room for their turn during the computer and stranger conditions. At the end of the experiment, we told both players that they were done, and didn't actually test the non-participant in the computer or stranger condition. For pairs who began in the computer or stranger condition, we had the non-participant wait outside of the experiment room until we were ready for the friend condition, and then released them both at the end of the experiment, again not testing the non-participant in the computer or stranger condition. For the duration of the experiment, both players thought that they were being tested, and it wasn't until the end of the experiment that one player realized that he would only play in one condition rather than three. Both participants received equal compensation of an Electronic Arts game of their choice to thank them for their participation.

Participants played NHL 2003 by EA Sports in both conditions (see Figure 12 for a screen shot). Six of the pairs were very experienced or somewhat experienced with the game, three pairs were neutral in their experience, while the other three pairs were somewhat inexperienced with the game.

Each play condition consisted of one 5-minute period of hockey. The game settings were kept consistent within each pair during the course of the experiment. All players used the Dallas Stars and the Philadelphia Flyers as the competing teams, as these two teams were comparable in the 2003 version of the game. All players used the overhead camera angle, and the home and away teams were kept consistent. This was

to ensure that any differences observed within subjects could be attributed to the change in play setting, and not to the change in game settings, camera angle, or direction of play. The only difference between pairs was that experienced pairs played all conditions in a higher difficulty setting than non-experienced players.

8.1.3 Experimental Setting and Protocol

The experiment was conducted in an office at Simon Fraser University. NHL 2003 was played on a Sony PS2, and viewed on a 36" television. A camera captured both of the players, their facial expressions and their use of the controller. All audio was captured with a boundary microphone. The game output, the camera recording, and the screen containing the physiological data were synchronized into a single quadrant video display, recorded onto tape, and digitized (see Figure 31) along with the audio from the game and the audio from the boundary microphone. The experimental setup was similar to the setup of Experiment Two, and a diagram of the complete experimental setup can be seen in Figure 20).

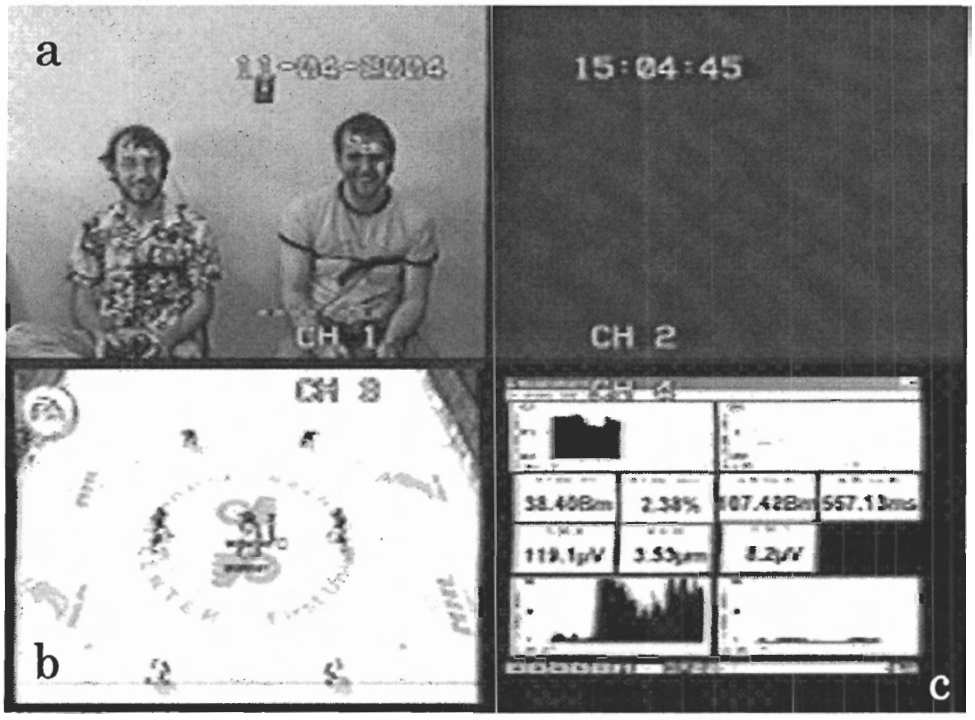


Figure 31: Quadrant display: a) camera feed of the participants, b) screen capture of the game, c) screen capture of the biometrics, audio of the game, and audio of the participants' comments.

Physiological data were gathered using the ProComp Infiniti system and sensors (see Figure 14), and BioGraph Software from Thought Technologies. Based on previous literature, we chose to collect galvanic skin response (GSR), electrocardiography (EKG), electromyography of the face ($EMG_{smiling}$ and $EMG_{frowning}$), and respiration. Heart rate (HR) was computed from the EKG signal, while respiration amplitude (RespAmp) and respiration rate (RespRate) were computed from the raw respiration data. We did not collect blood volume pulse data (BVP) because the sensing technology used on the finger is extremely sensitive to movement artifacts. As our subjects were operating a game controller, it wasn't possible to constrain their movements. Although we collected respiration data, we did not use respiration in this study. Respiration is most accurately measured by gas exchange in the lungs, but the

sensor technology inhibits talking and moving [123]. Instead, chest cavity expansion can be used to capture breathing activity using either a Hall effect sensor, strain gauge, or a stretch sensor, which produces much noisier data. The noise from using a stretch sensor is amplified in the computed respiration rate and amplitude. Although this noise can be treated when using the mean respiration rate or amplitude, we examine the entire time series in this experiment. As such, although we collected respiration data, it wasn't feasible to include respiration in our model.

Because we collected EMG in two locations on the face, we needed to gather five physiological signals for each participant. The ProComp Infiniti system that we used to collect the data (see Figure 14) only allows for eight inputs. As a result, we only collected physiological data for the participant, not for the friend or stranger. To maintain the perception that both players were participants in the experiment, we treated both players as if their physiological signals were being collected. We fitted both players with sensors, "tested" the sensor placement to ensure that the signals were good, and plugged the extra sensors into ports on the back of the unit.

Upon arriving, participants signed a consent form (see Appendix 4). They were then fitted with the physiological sensors. Before each experimental condition, participants rested for 5 minutes. Because the participants were required to rest in the same room before playing each other, they wore headphones and listened to a CD containing nature sounds. They also listened to the CD when resting alone to maintain consistency. The resting period allowed the physiological measures to return to baseline levels prior to each condition. Experiment One showed that the act of filling

out the questionnaires and communicating with the experimenter altered the physiological signals. The resting periods corrected for these effects. In order to utilize the resting periods as baseline controls, we would need much longer rest periods, and ensure that the nature sounds were indeed restful. We wanted to create an environment that was as natural as possible, and extended periods of rest in between play conditions did not fit with this approach.

After each condition, the participants filled out a condition questionnaire. The condition questionnaire contained their participant ID, the condition name, the level of play, and the final score (see Appendix 10). We also had subjects rate the condition using a Likert Scale. They were asked to consider the statement, "This condition was boring", rating their agreement on a 5-point scale with 1 corresponding to "Strongly Disagree" and 5 corresponding to "Strongly Agree". The same technique was used to rate how challenging, easy, engaging, exciting, frustrating, and fun that particular condition was. The questionnaire was filled out online using a laptop computer. Experiment One revealed that the physiological measurements for all participants reacted strongly to the interview process between each condition. We don't know what caused this effect but feel that the act of speaking and answering questions may have contributed. As a result, we chose to have participants fill out questionnaires online and then rest again for 5 minutes. After completing the experiment, subjects completed a post-experiment questionnaire using a laptop computer (see Appendix 11). We asked them to decide in retrospect which condition was most enjoyable, most fun, most exciting, and most challenging. They were also asked which condition they would choose to play in, given the choice to play against a co-located friend, against a

co-located stranger, or against the computer. Discussion of their answers was encouraged.

8.1.4 Data Analyses

The subjective data from the condition and post-experiment questionnaires were analyzed using non-parametric statistical techniques. In terms of the physiological data, EKG data were collected at 256 Hz, while GSR, respiration, and EMG were collected at 32 Hz. HR was computed at 4 Hz. Physiological data for each rest period and each condition were exported into a file. Noisy EKG data may produce heart rate (HR) data where two beats have been counted in a sampling interval or one beat has been counted in two sampling intervals. We inspected the HR data and corrected these erroneous samples, as described in section 6.4. In addition, HR data were interpolated since HR was sampled at a lower frequency than the EMG or GSR signals. After interpolation, HR was smoothed using a 4 frame moving average window.

Each EMG signal was smoothed with a moving average window of 4 frames (0.125 seconds) [39], while GSR was filtered using a 5-second window [10]. We then normalized each signal into a percentage between 0 and 100. There are very large individual differences associated with physiological data, and normalizing the data is necessary to perform a group analysis. We transformed each sample into a percentage of the span for that particular signal, for each participant across all three conditions. Using GSR as an example, a global minimum and maximum GSR were obtained for each participant using all three conditions and the rest period, and the same global values were used for normalizing within each condition.

$$\text{Normalized GSR(i)} = \left(\frac{\text{GSR(i)} - \text{GSRmin}}{\text{GSRmax} - \text{GSRmin}} \right) \times 100$$

The same method was used to normalize the $EMG_{smiling}$, $EMG_{frowning}$, and HR data.

8.2 Fuzzy Logic

We used normalized GSR, HR, $EMG_{smiling}$, and $EMG_{frowning}$ signals as inputs to a fuzzy logic model. To generate values for user emotion, we modeled the data in two parts. First, we computed arousal and valence values from the normalized physiological signals, then used these arousal and valence values to generate emotion values for boredom, challenge, excitement, frustration, and fun.

Fuzzy logic mimics human control logic in that it uses an imprecise but descriptive language to deal with input data, much like a human operator [20]. Fuzzy logic systems address the imprecision of the input and output variables by defining them with fuzzy numbers and fuzzy sets that are expressed in linguistic terms (e.g., cold, warm, hot) [131]. Simple, plain-language IF/THEN rules are used to describe the desired system response in terms of the linguistic variables, rather than through complex mathematical formulas [57].

If we wanted to classify temperatures as cold, warm, or hot, classical sets would require hard boundaries and binary memberships. For example, a set of all warm temperatures *between 15°C and 35°C* would not include a temperature of 35.01°C. Fuzzy sets use linguistic definitions and could include temperatures around the boundaries. Binary memberships still exist, with 25°C being a full member of the

warm set and 50°C existing fully outside of the warm set. But fuzzy sets allow for partial membership around the boundaries. Figure 32a shows how a classical set has firm boundaries and binary memberships for classifying temperatures, whereas fuzzy sets allow for partial membership. In fuzzy sets, the membership functions transform the membership of a specific temperature into a degree of membership in the set. Membership functions can take a number of shapes; however, triangular and trapezoidal membership functions are the most common [131]. The trapezoidal membership function in Figure 32b specifies that temperatures between 20°C and 30°C have full membership in the warm set, while temperatures from 15°C to 20°C and 30°C to 35°C have partial membership in the set. The temperatures that are closer to 30°C have a greater degree of membership than those that closer to 35°C. With fuzzy sets, the values that exist in the boundaries between sets can exist in both sets. In our example, 35°C has partial membership in the warm set and partial membership in the hot set, and has an equal degree of membership in both sets. The value of 33°C also has membership in both the warm and hot sets, but has a greater degree of membership in the warm set than in the hot set.

Fuzzy logic can easily represent continuous processes that are not easily broken into discrete segments, when the change of state from one linguistically-defined level to the next is not clear [20]. For example, there does not have to be a definitive point when a rising temperature moves from cold to warm. In general, fuzzy logic should be used when [20]:

1. one or more of the control variables are continuous;
2. when a mathematical model of the process does not exist;
3. when high ambient noise levels must be dealt with;
4. when an expert can identify the rules underlying the system behaviour and the fuzzy sets that represent that characteristics of each variable.

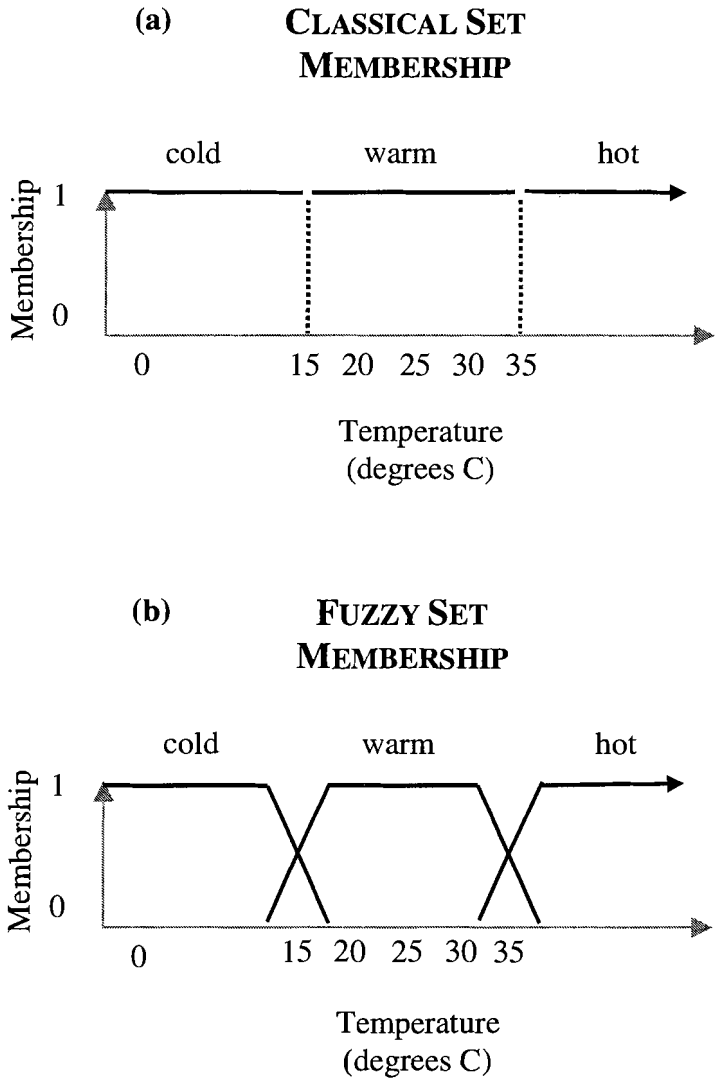


Figure 32: **A graphical representation of set membership for classifying temperature for both classical (a) and fuzzy (b) sets. Classical sets have firm boundaries and binary membership, whereas fuzzy sets allow for partial membership around the edges.**

The fuzzy logic system consists of inputs, outputs, membership functions, and rules. The membership function is a graphical representation of the magnitude of participation of each input [57]. It weights each input signal, defines overlap between the levels of input, and determines an output response. The IF/THEN rules use the input membership values as weighting factors to determine their influence on the fuzzy solution sets [20, 57]. Once the functions are inferred, scaled, and combined, they are defuzzified[8] into a solution variable (scalar output) [20, 57]. Membership functions can be different for each input and output response.

There are other machine learning methods available, including neural nets. Neural nets and fuzzy systems take opposite approaches to dealing with uncertainty [131]. Neural nets use precise inputs and outputs which are used to train a generic model, while in fuzzy systems, the inputs and outputs are fuzzy and their interrelationships take the form of well-defined IF/THEN rules [131]. One of the disadvantages of neural nets is that they need substantial data that cover the entire range over which the different variables are expected to change [131]. Our participants are generally happy; however, there could easily be moments when participants are bored or frustrated. We cannot guarantee that the complete span of any emotion will be covered by game playing.

Fuzzy logic systems are best used when variables are continuous [20], as with the physiological signals that we collect. We chose to use a fuzzy approach since there is

---

[8] We used the centroid method of defuzzification.

a strong theoretical basis for the transformation from input to output; an expert can use linguistic terms to describe this transformation; we have noisy input signals; and the physiological variables are continuous.

8.3 Modeling Arousal-Valence Space

The first stage was to transform the physiological signals into AV space (arousal-valence space). To generate the models, we used half of the participants (one for each play condition order), reserving the other six participants for validation of the model. We randomly chose which participants were used to generate the model, and which were used for the validation of the model. To make use of the continuous nature of physiological data, we used the complete time series for each input. As such, we were able to generate a new time series of the participant's experience in AV space, rather than having only one data point for an entire condition (e.g. mean).

Our model to transform physiology to AV space had four inputs (GSR, HR, $EMG_{smiling}$, and $EMG_{frowning}$) and two outputs (arousal and valence) (see Figure 33). Inputs were normalized signals (0-100), while outputs were percentages of the possible maximum (0-100) value for arousal and valence.

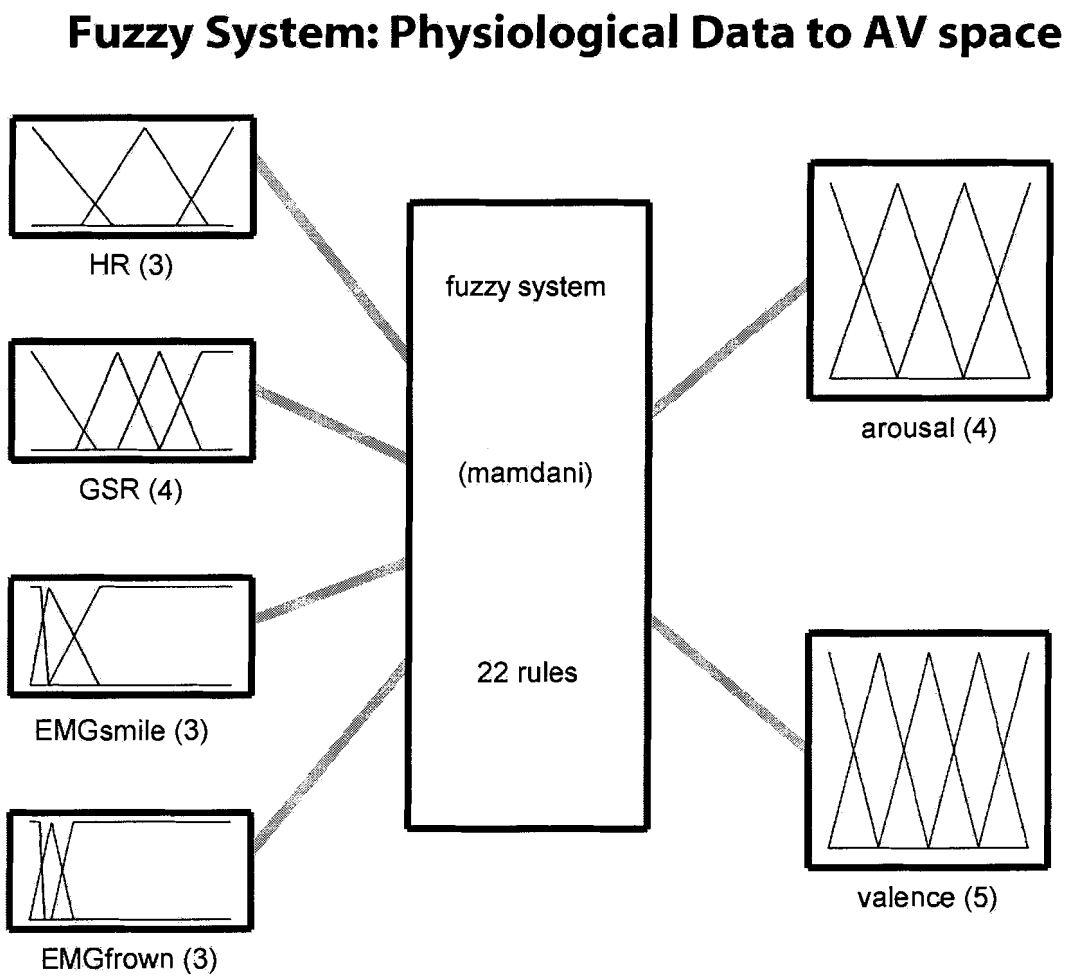


Figure 33: **Modeling arousal and valence from physiological data. The number of membership functions applied to that input or output follows the input / output labels. The system used 22 rules to transform the 4 inputs into the 2 outputs.**

8.3.1 Membership Functions

Membership functions were applied to the four physiological inputs and the two outputs. In terms of the inputs, the membership functions describe what defines a low, medium, or high value of the input. The fuzzy aspect comes in such that any value of the input doesn't *necessarily* belong to any one set (low, medium, or high), but there are areas of overlap between the levels. For example, a HR input value of 30% may fall in a "fuzzy" area where it could be considered a low or medium value of HR.

8.3.1.1 Input Data Histograms

For each input signal, the membership functions were generated using characteristics of that particular signal over the six participants and three conditions. For each of the input signals, there are a total of 147176 samples. We generated histograms for each input, with 1000 bins, in order to have approximately 150 samples per bin. These values were chosen to maximize the number of bins while maintaining statistical relevance, and to ensure the division of value didn't exceed the precision of measurement of the samples (see Figure 34: HR; Figure 35: GSR; Figure 36: $EMG_{smiling}$; Figure 37: $EMG_{frowning}$).

8.3.1.2 Derivation of the Membership Functions

Figure 38 through Figure 41 show how the membership functions were generated for each input signal, using the statistical characteristics of the histograms shown in the previous section. As seen in Figure 34, HR approaches a normal distribution, where 68.27% of the area under the curve is within one standard deviation of the mean, and 95.45% of the area is within two standard deviations. For HR, these characteristics

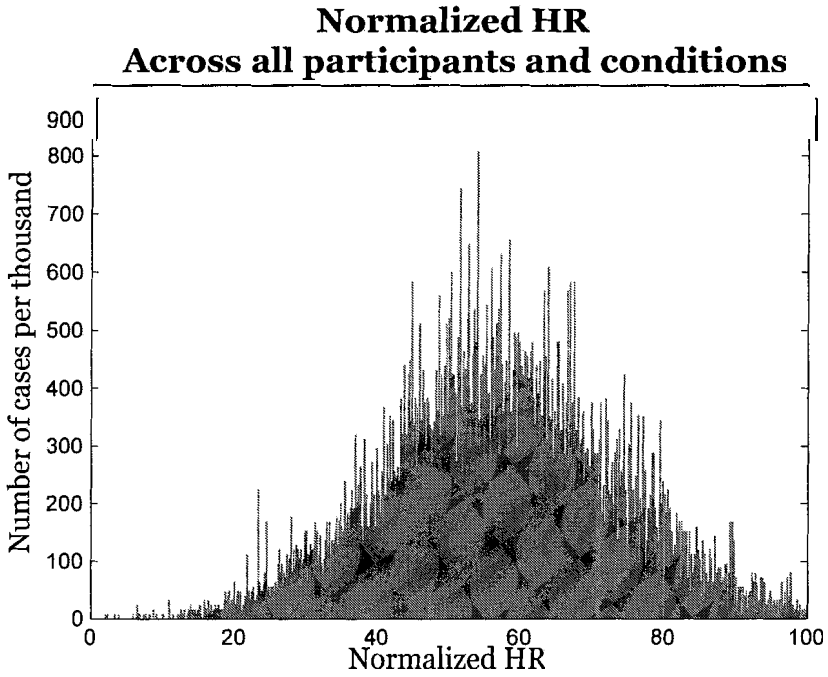


Figure 34: Histogram of normalized HR for all six participants across all three play conditions. HR approximates a normal distribution.

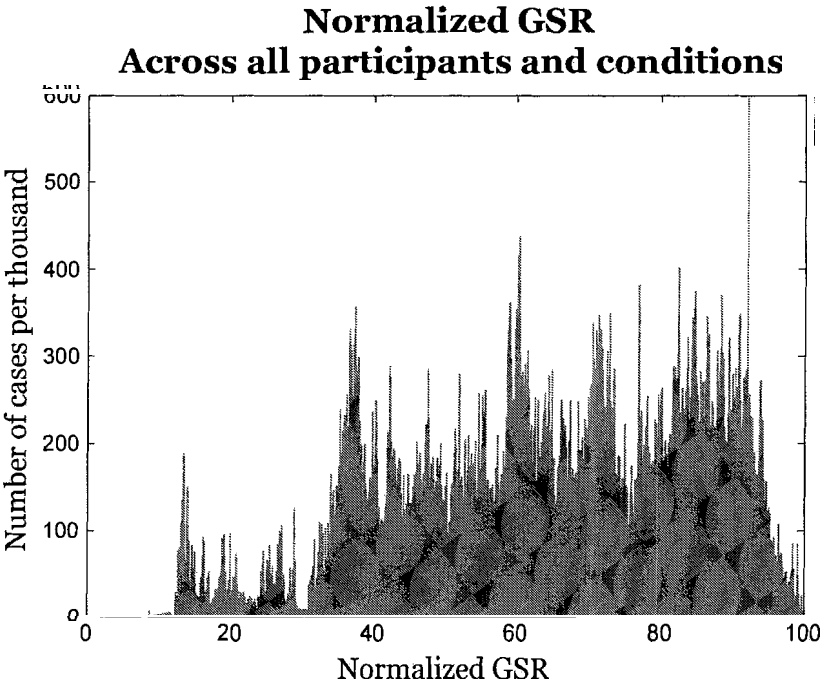


Figure 35: Histogram of normalized GSR for all six participants across all three play conditions. GSR is a multi-peaked non-normal distribution.

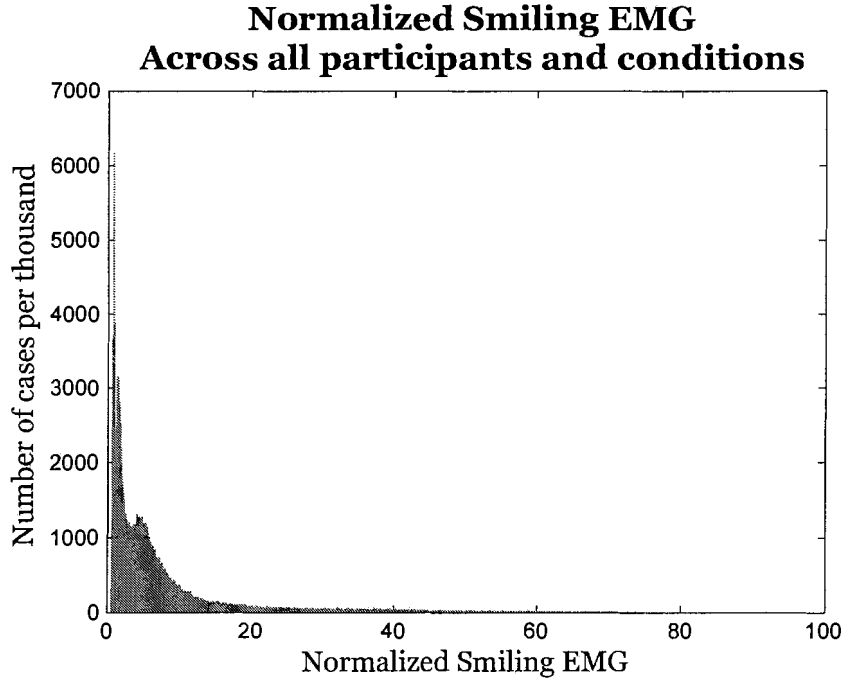


Figure 36: Histogram of normalized $EMG_{smiling}$ for all six participants across all three play conditions. $EMG_{smiling}$ approximates a lognormal distribution.

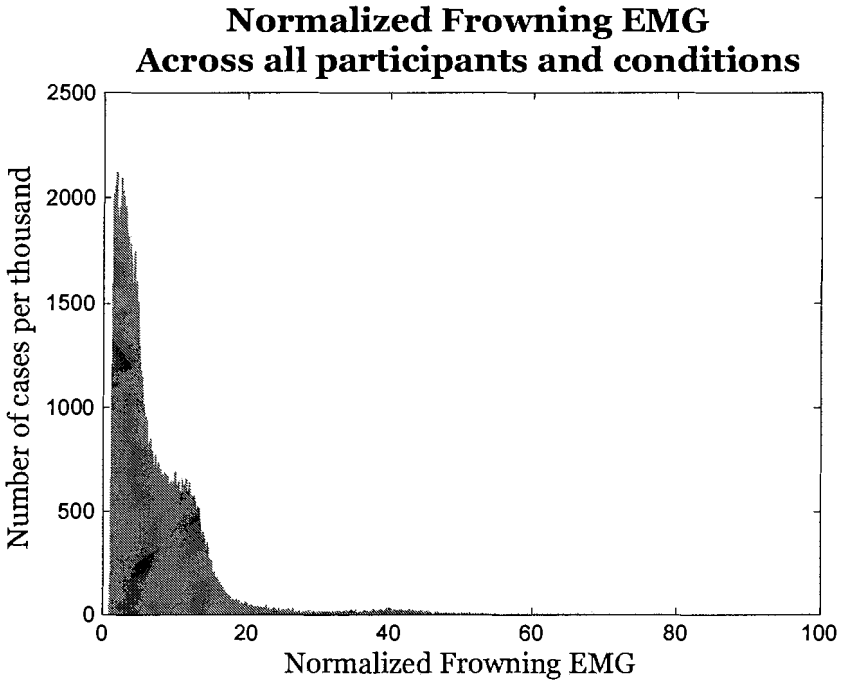


Figure 37: Histogram of normalized $EMG_{frowning}$ for all six participants across all three play conditions. $EMG_{frowning}$ approximates a lognormal distribution.

were used to define membership functions that suit the distribution of the input signal. Figure 38 shows how three membership functions describe low, medium, and high HR activity. The membership functions were all triangular, as seen in Figure 33.

Figure 35 shows how GSR was distributed across the entire span, although more activity occurred in the mid and high range. As the distribution of GSR contained multiple peaks, four membership functions were used: low, mid-low, mid-high, and high. The statistical characteristics of the signal were used to determine where the membership functions were positioned (see Figure 39). The membership functions were triangular and trapezoidal as seen in Figure 33.

Both $EMG_{smiling}$ and $EMG_{frowning}$ were clustered towards the low end of activation (see Figure 36 and Figure 37), approximating lognormal distributions. For both EMG signals, three membership functions were defined, representing low, medium, and high EMG activity. Due to the statistical characteristics of a lognormal distribution, the membership functions were clustered towards the low end of activation (see Figure 40 and Figure 41). The medium membership function was triangular, while the low and high membership functions were trapezoidal. The trapezoids were used to remove fuzziness from the extreme values of input.

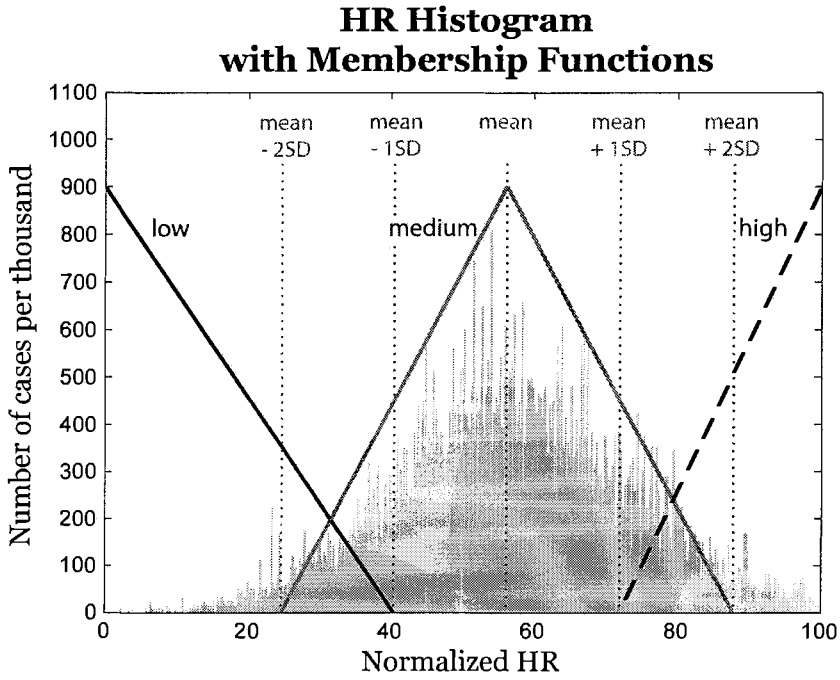


Figure 38: Histogram of HR with statistical characteristics and three membership functions superimposed.

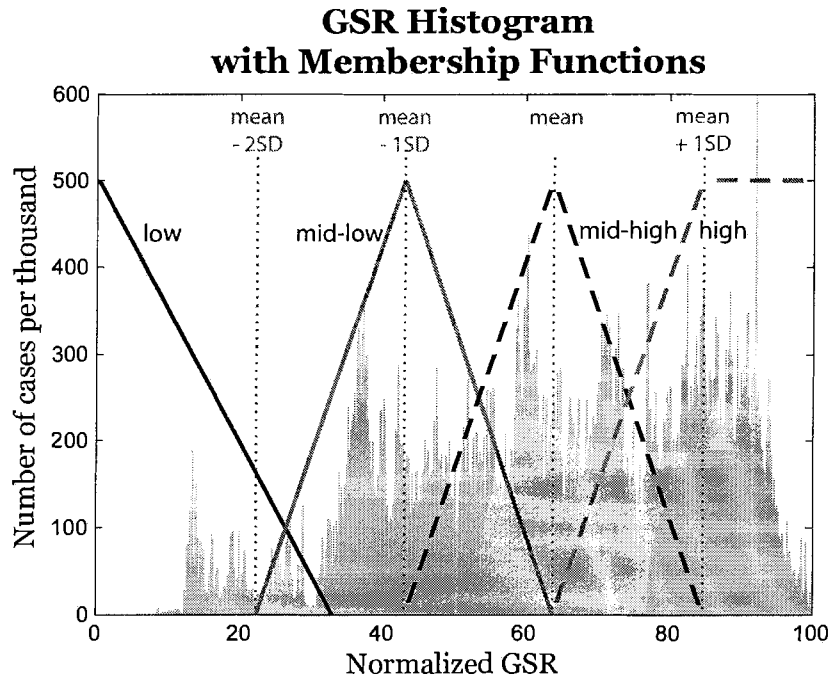


Figure 39: Histogram of GSR with statistical characteristics and four membership functions superimposed.

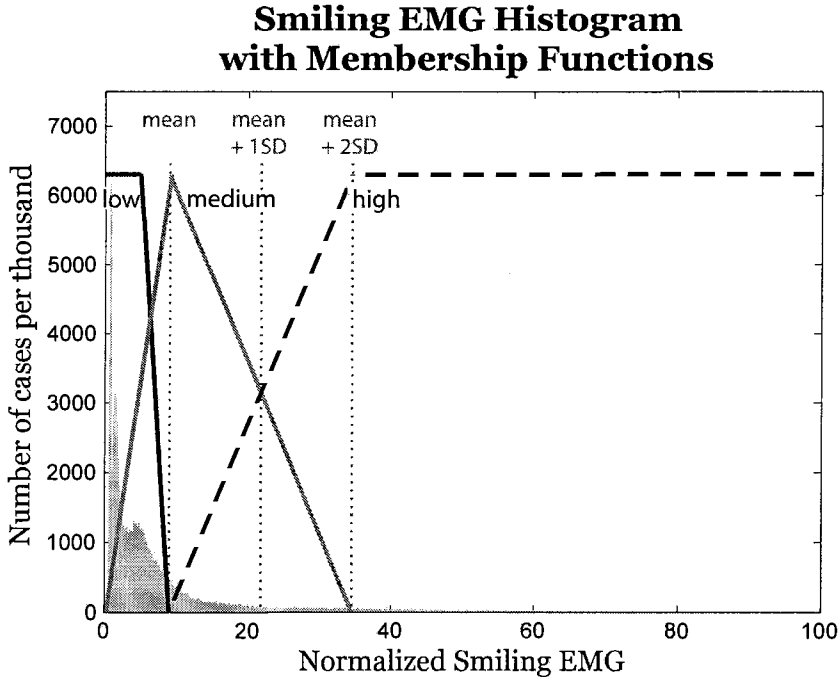


Figure 40: Histogram of $EMG_{smiling}$ with statistical characteristics and three membership functions superimposed.

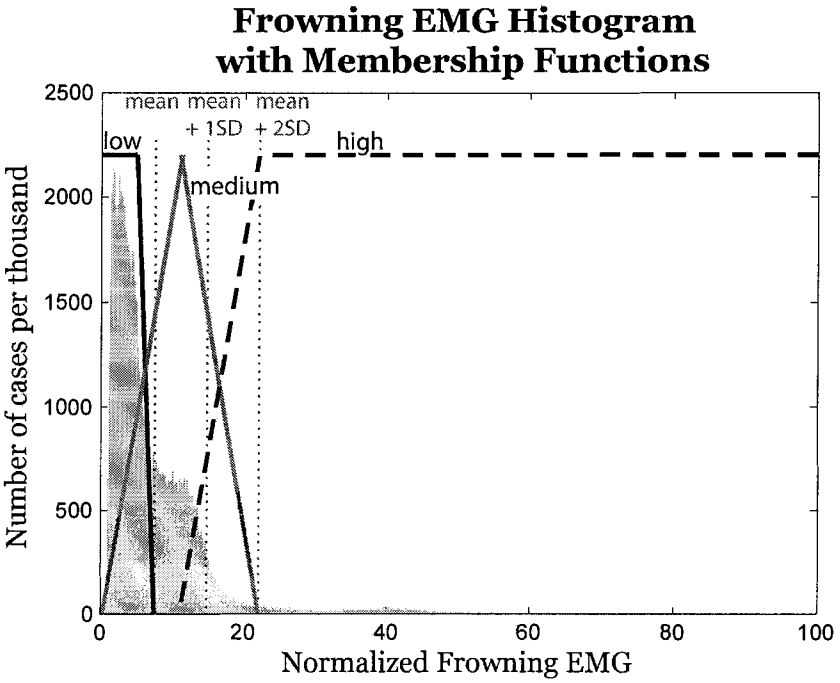


Figure 41: Histogram of $EMG_{frowning}$ with statistical characteristics and three membership functions superimposed.

Membership functions for the two outputs (arousal and valence) were distributed evenly across the entire spectrum. Arousal was defined with four memberships: low, mid-low, mid-high, and high. Valence was described by five memberships: very low, low, neutral, high, and very high. The neutral membership was introduced to accommodate the large percentage of smiling and frowning activity that occurred at less than 5% of total activation. The output membership functions were all triangular as seen in Figure 33.

8.3.2 Rules

The 22 rules were grounded in the theory of how the physiological signals relate to the psychological concepts of arousal and valence. Arousal was generated from GSR and HR, while valence was generated from $EMG_{smiling}$, $EMG_{frowning}$, and HR.

GSR correlates with arousal, and increasing GSR was mapped to increasing arousal. The extreme high and low levels of GSR were modulated by HR data; if HR contradicted GSR, arousal was altered, otherwise arousal was maintained. Figure 42 shows how GSR and HR combine through the defined rules and membership functions to generate arousal.

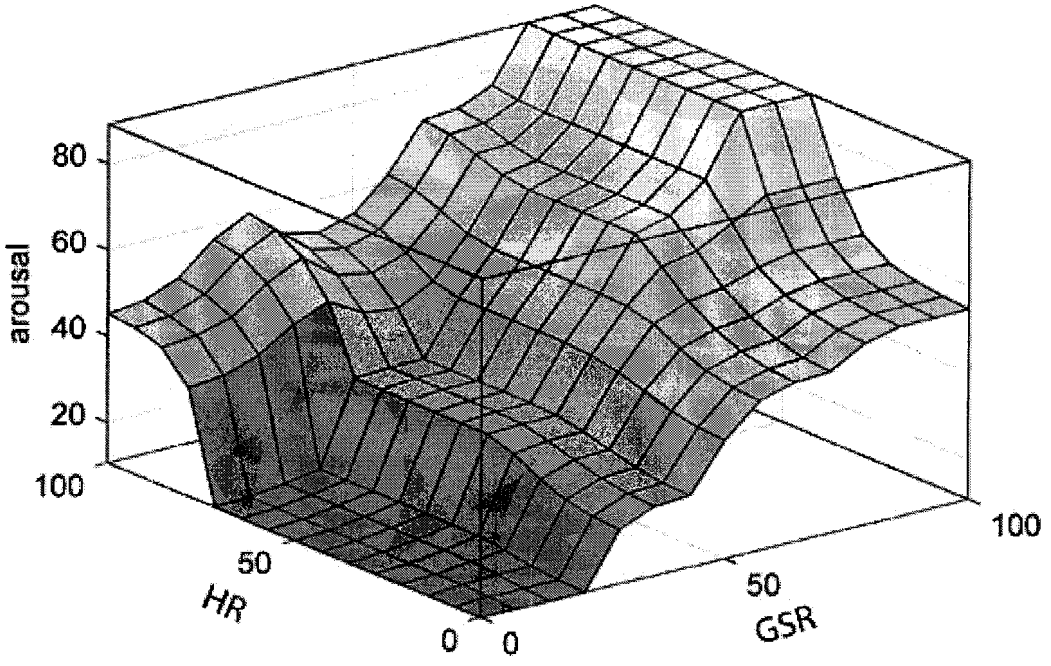


Figure 42: GSR and HR combine to generate arousal. GSR has more of an impact on arousal; however, arousal is modulated by HR when HR and GSR are contradictory. This is reflected in the 'wings' on the arousal surface.

Since smiling activity reflects positive emotions, and frowning activity represents negative emotions, valence generally increased with increasing levels of $EMG_{smiling}$, and decreased with increasing levels of $EMG_{frowning}$. Figure 43 shows how $EMG_{smiling}$ and $EMG_{frowning}$ combine through the rules and membership functions to generate valence. Because the majority of the activation for both EMG signals occurred at less than 5%, (neutral facial expression) we would expect valence to be neutral most of the time. In addition, when $EMG_{smiling}$ and $EMG_{frowning}$ were both high, the valence output resolved to a neutral state. This type of activation would occur when participants were making a face other than smiling or frowning, and did not occur very often. When both EMG signals are low, EMG does not provide enough information to predict

valence. As a result, we used HR to modulate these occurrences (see rules 18 and 19 below). HR tends to increase with positive affect [97, 150], so when we were unable to distinguish valence for EMG alone, we used high HR values to move valence from neutral to high, and low HR values to move valence from neutral to low. The 22 rules are presented in Appendix 12.

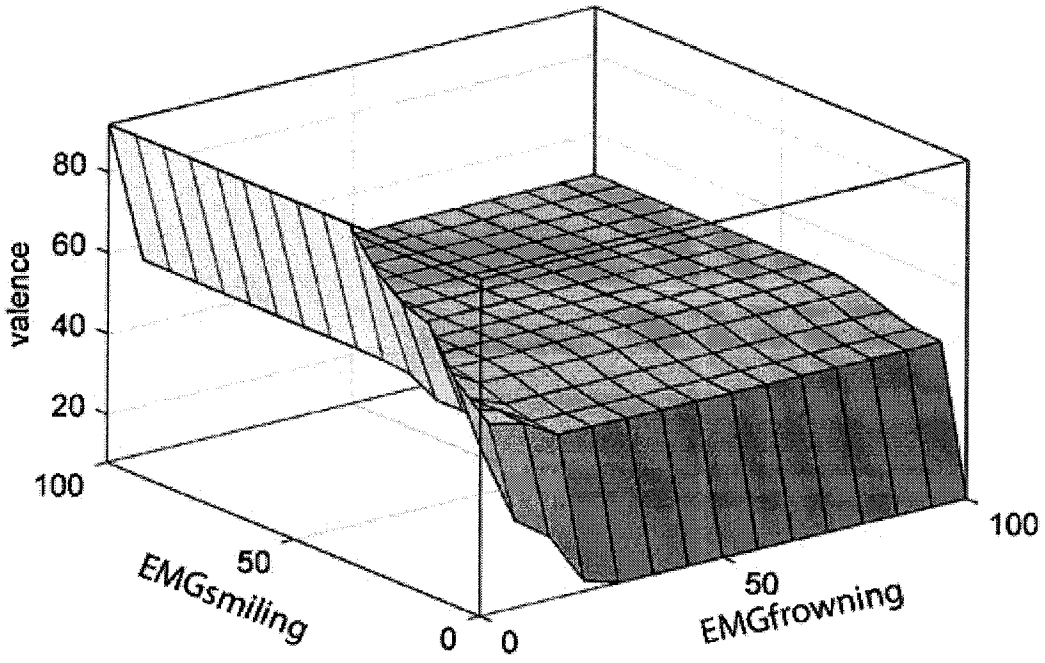


Figure 43: **$EMG_{smiling}$ and $EMG_{frowning}$ are converted into valence. Since the majority of the activation for both EMG signals occurred at less than 5%, (neutral facial expression) we would expect valence to be neutral most of the time. In addition, when $EMG_{smiling}$ and $EMG_{frowning}$ were both high, the valence output resolved to a neutral state.**

8.3.3 Fuzzy Approach Results

Experiment Two revealed that GSR and EMG_{jaw} were higher when playing against a friend, over playing against a computer. We would expect that arousal and valence would be higher when playing against a friend, over playing against the computer. To examine whether our model is achieving the predicted results, we looked at the mean values of arousal and valence across the play conditions.

The mean results are shown in Table 15. A repeated measures ANOVA shows that there was a significant difference in valence between the three play conditions. Post-hoc analysis revealed that valence was higher when playing against a friend than when playing against the computer ($p = .005$). There was no significant difference in arousal between the conditions, although mean arousal was greater when playing against a friend over playing against a computer.

Table 15: Mean arousal and valence values from the fuzzy approach. There was a difference in valence between conditions, but not in arousal.

	Playing against computer		**Playing against friend**		**Playing against stranger**		**Difference between conditions**		
	Mean	*St. Dev.*	*Mean*	*St. Dev.*	*Mean*	*St. Dev.*	*F*	*p*	η^2
arousal	66.2	23.5	69.7	11.9	71.9	31.2	0.09	.919	.02
valence	**65.5**	**7.4**	**71.9**	**7.1**	**68.1**	**6.2**	**5.70**	**.022**	**.53**

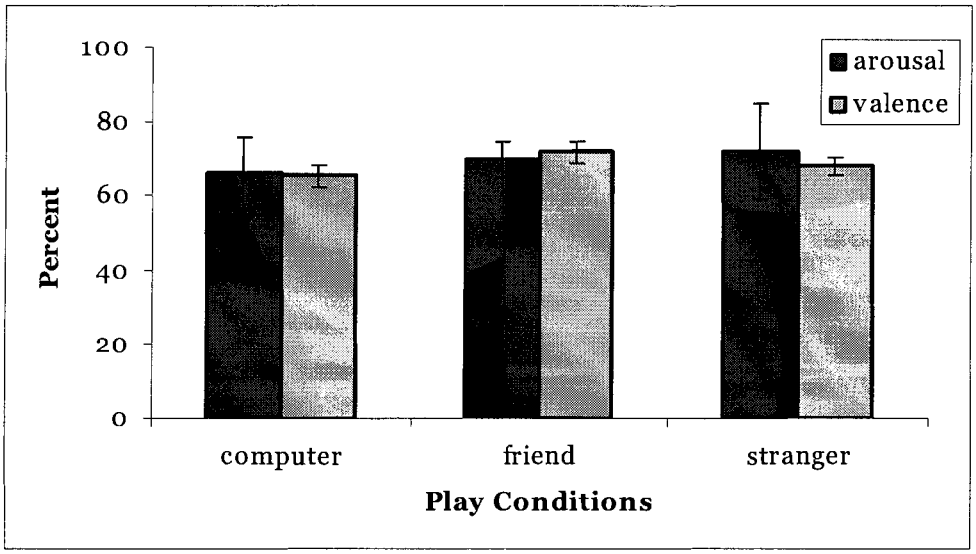


Figure 44: Mean results of arousal and valence (±SE) from the fuzzy approach, separated by play condition.

8.3.4 Manual Approach

We also used a manual approach to calculate arousal and valence for each sample. The manual approach was implemented in order to confirm that the output from the fuzzy logic model was on track. For the manual calculations, we used the normalized GSR signal as the arousal metric since GSR is a linear correlate to arousal. For valence, we took normalized $EMG_{smiling}$, and subtracted normalized $EMG_{frowning}$, and re-normalized to generate a number between 0 and 100.

Table 16: Mean arousal and valence values from the manual approach. There was a difference in valence between conditions, but not in arousal.

	Playing against computer		Playing against friend		Playing against stranger		Difference between conditions		
	Mean	*St. Dev.*	*Mean*	*St. Dev.*	*Mean*	*St. Dev.*	*F*	*p*	η^2
arousal	63.1	21.3	64.7	10.9	66.4	28.1	.97	.967	.01
valence	**47.2**	**2.5**	**52.7**	**2.6**	**49.0**	**2.3**	**21.2**	**.001**	**.81**

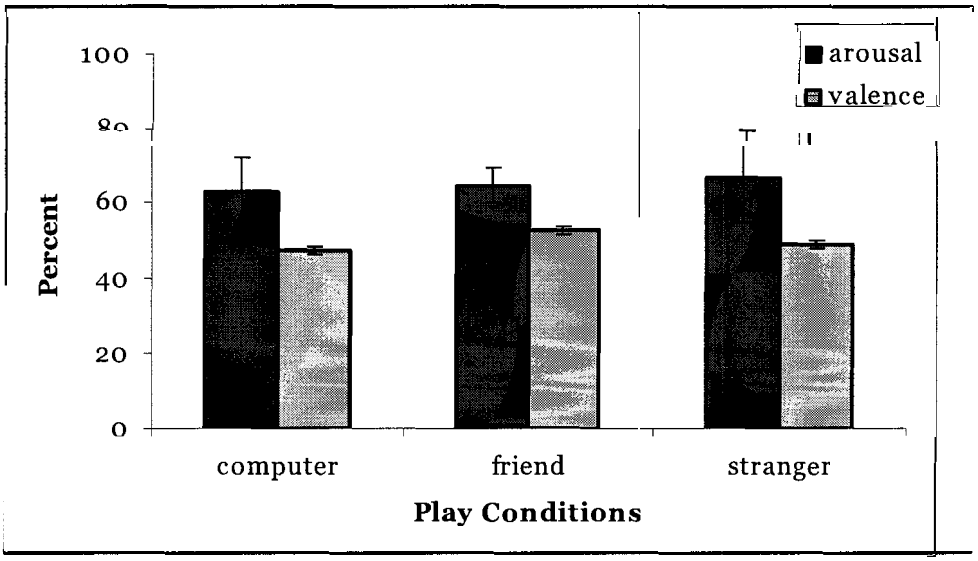


Figure 45: Mean results of arousal and valence (±SE) from the manual approach, separated by play condition.

The mean results are shown in Table 16. A repeated measures ANOVA shows that there was a significant difference in valence between the three play conditions. Post-hoc analysis revealed that valence was higher when playing against a friend than when playing against the computer ($p = .001$) or a stranger ($p = .005$). There was no difference in arousal between conditions.

8.3.5 Comparing Fuzzy and Manual Results

We wanted to compare the arousal and valence results from the fuzzy model to the results from a manual approach using a distance metric. As such, we took the absolute difference between the fuzzy result and the manual result for each value for arousal and valence for all six participants, in all three conditions. The mean differences and maximum differences for each condition are shown in Table 17, while Figure 46 through Figure 51 show histograms of the total differences in arousal and valence for

each condition. When averaged for each condition, the mean differences between the fuzzy and manual approach were between 3% and 6% for both arousal and valence. The maximum difference between the fuzzy and manual approaches for both arousal and valence occurred in the friend condition (arousal = 20.4% and valence = 41.8%).

In all, the fuzzy approach performs in a very similar manner to the manual approach. Differences were computed for every sample in the time series, (a total of 147176 samples), yet average differences were only on the order of 5%, and maximum differences were always less than 50%.

We used a repeated measures ANOVA to see if the manual and fuzzy approaches were more or less comparable in each play condition. There was a significant difference in mean valence difference (see Table 17). Post hoc analysis revealed that for valence, the manual and fuzzy approaches were more similar in the stranger ($p = .010$) and computer condition ($p = .035$), than in the friend condition.

Table 17: Mean differences between the manual approach and the fuzzy approach, separated by condition. Mean valence difference was higher in the friend condition than in the computer or stranger condition.

	Playing against computer		**Playing against friend**		**Playing against stranger**		**Difference between conditions**		
	Mean	*St. Dev.*	*Mean*	*St. Dev.*	*Mean*	*St. Dev.*	*F*	*p*	η^2
mean arousal diff. (%)	5.3	3.4	3.6	1.6	3.4	0.6	1.29	.316	.21
mean valence diff. (%)	**3.9**	**2.3**	**5.5**	**1.6**	**3.7**	**1.9**	**9.83**	**.004**	**.66**
max arousal diff. (%)	19.4	10.2	20.4	9.9	16.6	7.0	0.39	.685	.07
max valence diff. (%)	26.6	9.6	41.8	8.4	30.3	13.4	3.27	.081	.40

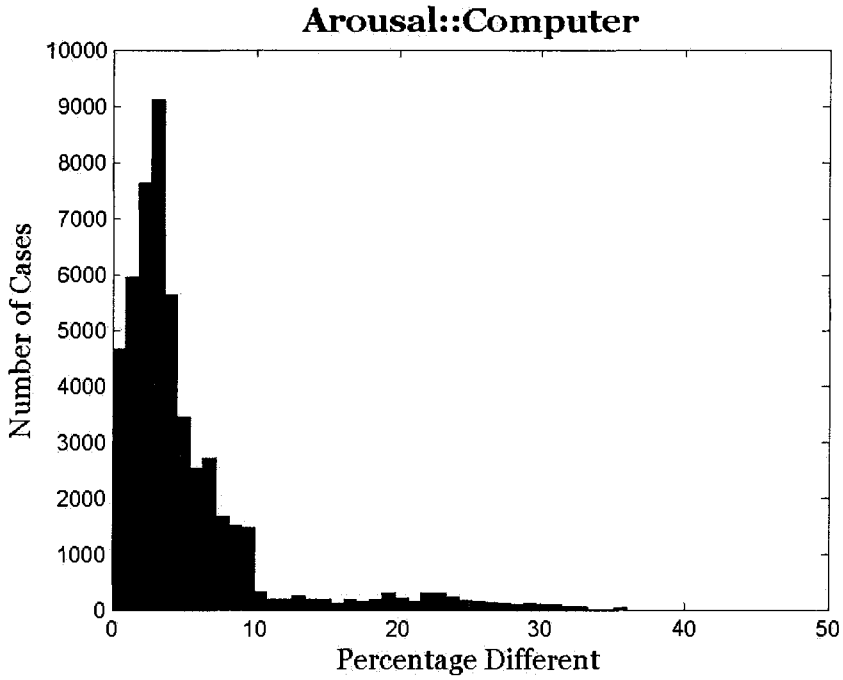


Figure 46: **A histogram reveals the total differences between the fuzzy and manual approaches for arousal in the computer condition. The majority of the samples were less than 5% different.**

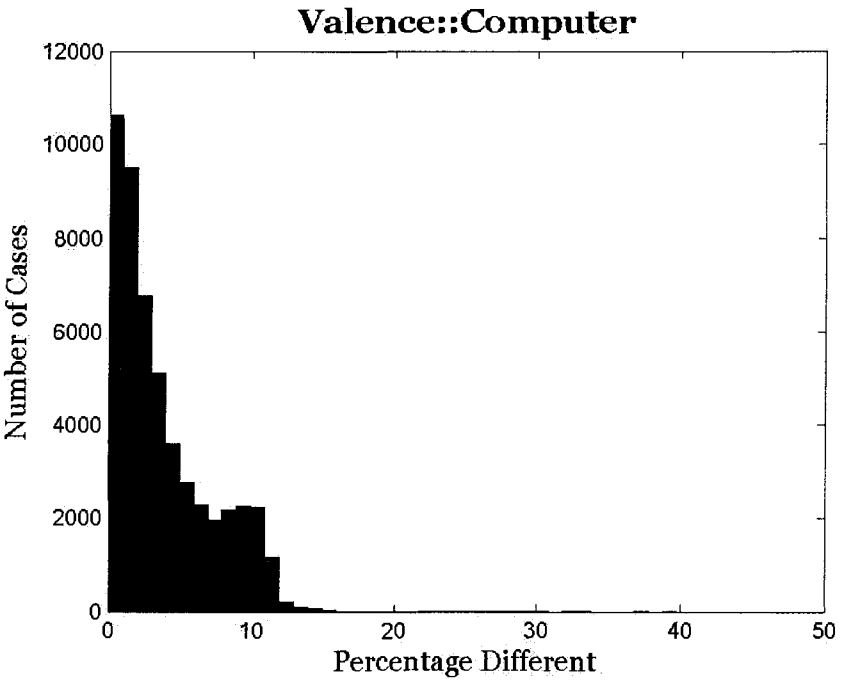


Figure 47: **A histogram reveals the total differences between the fuzzy and manual approaches for valence in the computer condition. The majority of the samples were less than 5% different.**

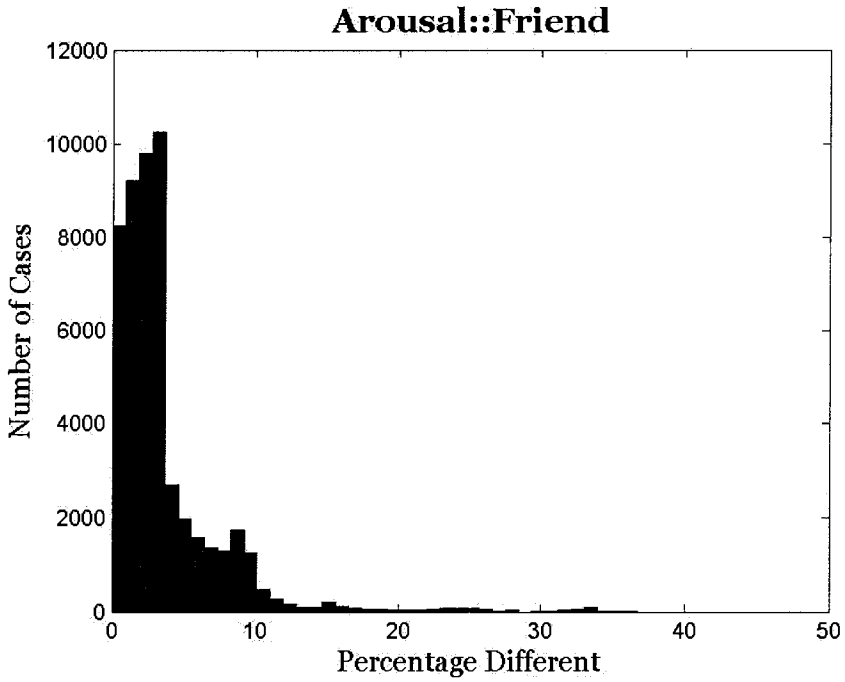


Figure 48: **A histogram reveals the total differences between the fuzzy and manual approaches for arousal in the friend condition. The majority of the samples were less than 5% different.**

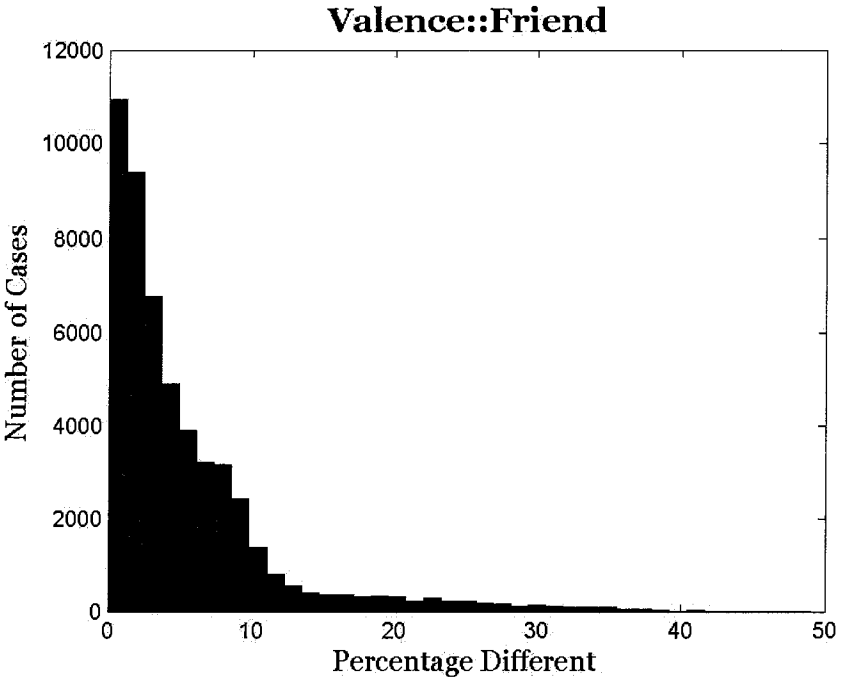


Figure 49: **A histogram reveals the total differences between the fuzzy and manual approaches for valence in the friend condition. The majority of the samples were less than 5% different.**

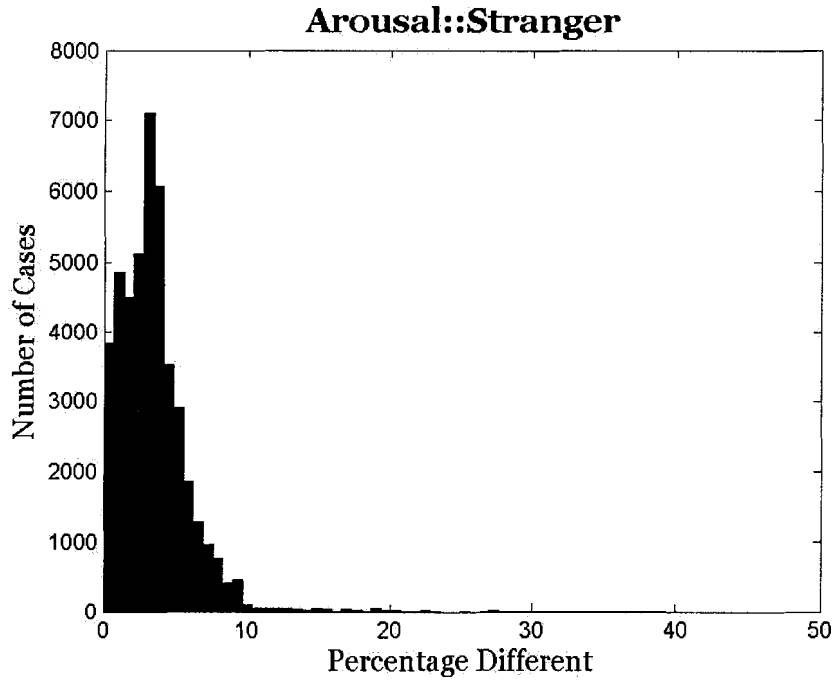


Figure 50: **A histogram reveals the total differences between the fuzzy and manual approaches for arousal in the stranger condition. The majority of the samples were less than 5% different.**

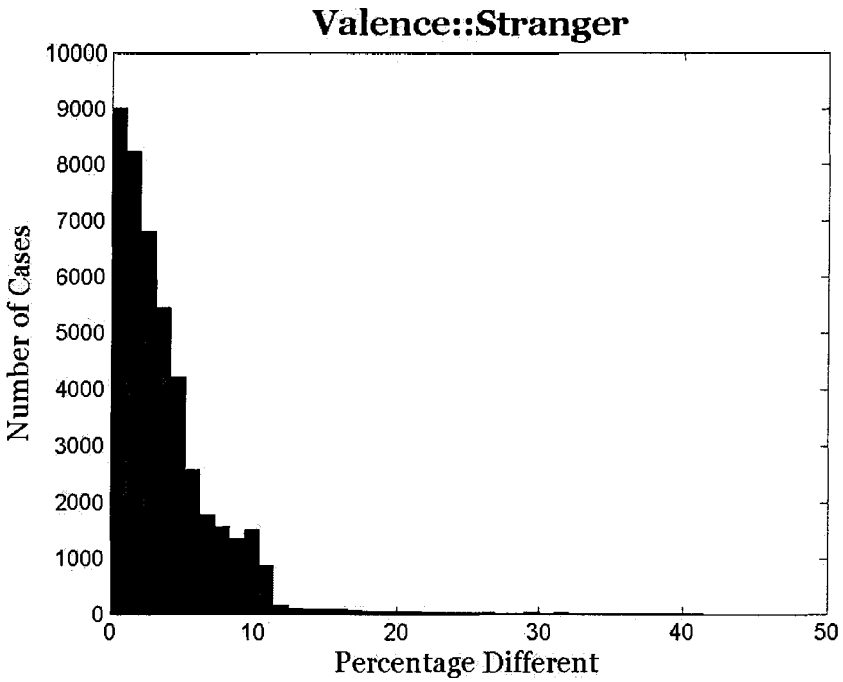


Figure 51: **A histogram reveals the total differences between the fuzzy and manual approaches for valence in the stranger condition. The majority of the samples were less than 5% different.**

8.3.5.1 AV-Space Graphs

The fuzzy and manual approaches reveal fairly similar results. In order to visualize how the two approaches differ, we generated graphs of a participant's experience in AV space over time. Traditionally, the affect grid [114] asks participants to mark an X to describe their experience in AV space. Since our approach is continuous, it is important to visualize their experience as it changed over time.

Appendix 14 shows all of the participants' experiences as graphed in AV space. In general, we noticed that the manual approach tends to place activity in the extreme areas of AV space. Figure 52 and Figure 53 show Participant 16's experience in AV space when playing against a friend. The manual approach (Figure 52) reaches the extreme positive values of both arousal and valence, whereas the fuzzy approach (Figure 53) is less reactionary, and more moderate.

The manual approach is also more reactive to participants' facial expressions. For example, when a participant smiles, their valence increases instantly to the maximum value, whereas the fuzzy approach is a bit more moderate in evaluating valence. Figure 54 and Figure 55 show the AV experience for Participant 16 playing against the computer. The manual approach (Figure 54) seems to use the neutral state as a 'home base'. Valence is generally neutral, but sometimes increases and subsequently returns to the neutral state. In contrast, the fuzzy approach (Figure 55) is much less volatile and there is more continuity in valence throughout the experience.

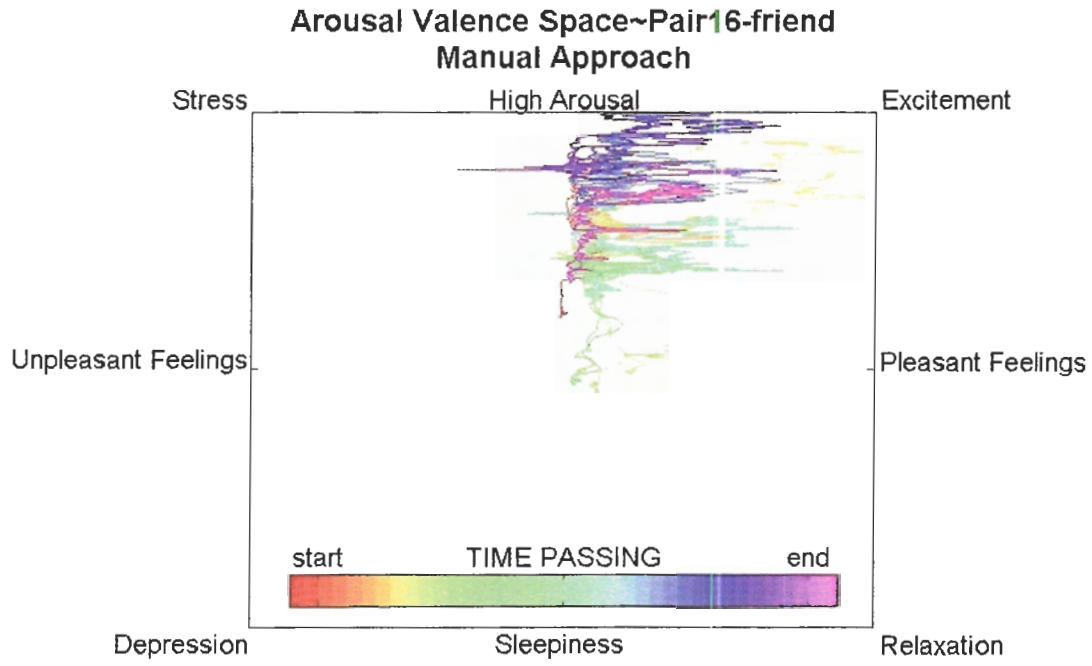


Figure 52: The experience of Participant 16, in AV space while playing against a friend. This graph is generated using the manual approach.

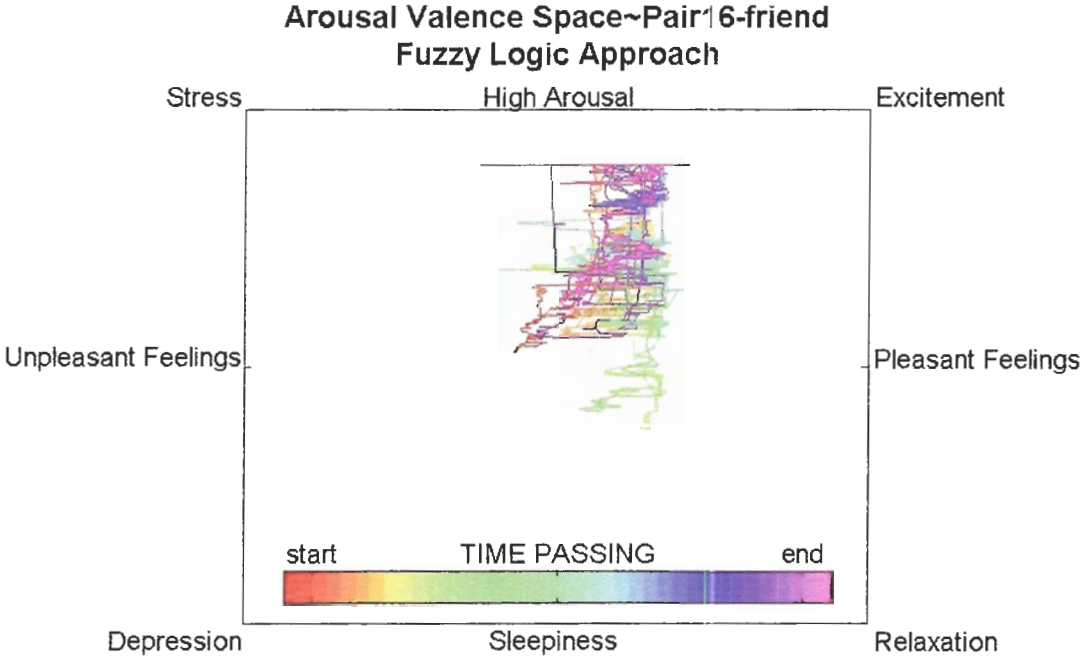


Figure 53: The experience of Participant 16, in AV space while playing against a friend. This graph is generated using the fuzzy approach.

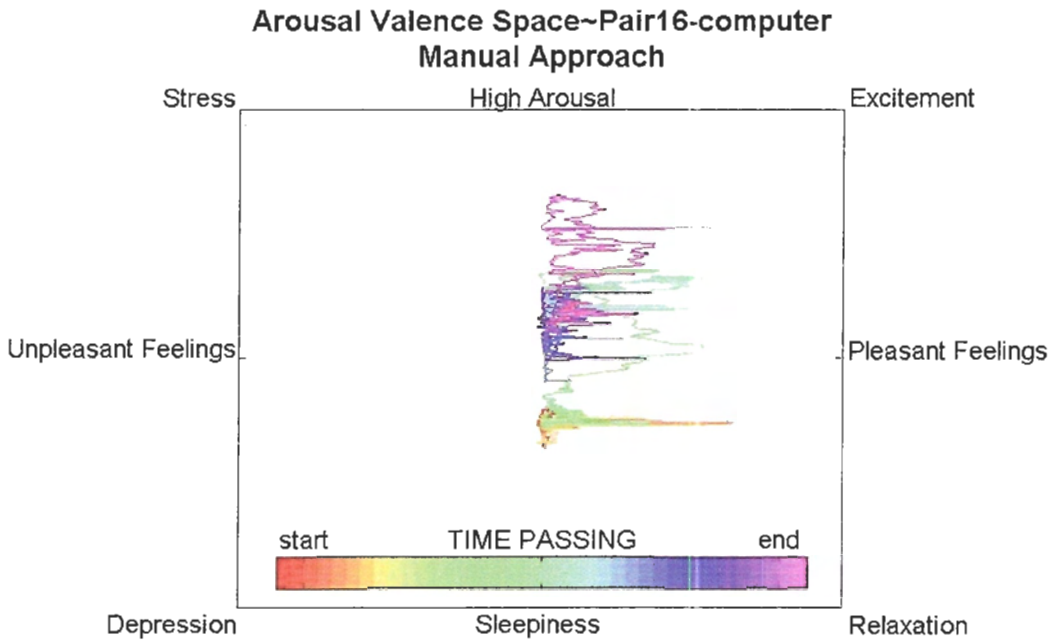


Figure 54: The experience of Participant 16, in AV space while playing against the computer. This graph is generated using the manual approach.

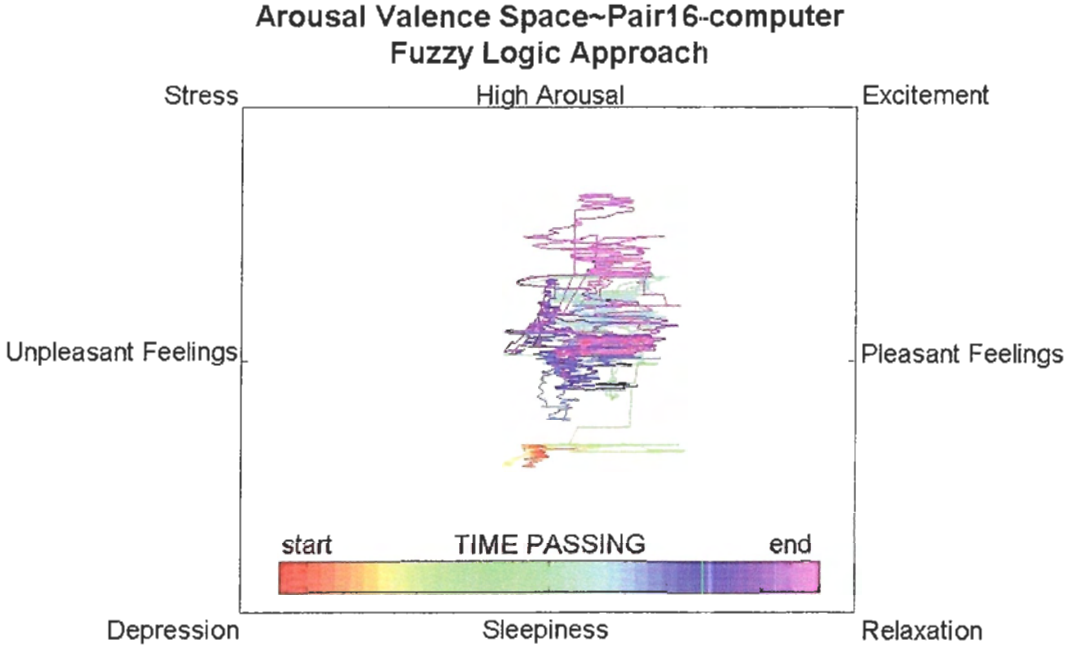


Figure 55: The experience of Participant 16, in AV space while playing against the computer. This graph is generated using the fuzzy approach.

8.3.6 Issues with Modeling Arousal and Valence

Although our AV space model is based in a theoretical understanding of the psychophysiology signals, there are some outstanding implementation issues involving the scaling of the arousal and valence axes. Our data successfully shows arousal and valence changing over time; however, the absolute positioning of this experience in AV space is difficult. In order to determine maximum arousal and valence, we used the minimum and maximum values from the all three play conditions and the rest period. We determined the baseline arousal and valence values to the best of our ability, given the available data; however, the available data may not have contained accurate baseline values.

A better approach to scaling the arousal and valence axes would have been to use the IAPS [64] to calculate baselines for participants' arousal and valence. Presenting pictures from the IAPS data set, and measuring a subject's responses could provide accurate scaling information that we could use to position that subject's game-playing experience in AV space. Although informative, this process would be riddled with logistic problems since GSR is not consistent across experimental sessions [10]. Baselining a participant's GSR response on one day might not apply to the following day or week. Using a variety of baselines and dynamically adjusting for the day-to-day variations would be a feasible approach, requiring additional research.

8.4 Modeling Emotion from AV Space

The second phase of the emotion model is to use the arousal and valence information to model different emotions. To make the most of the rich, continuous physiological

data, we modeled the entire AV space time series, creating continuous metrics of emotional experience. Five emotions were modeled: boredom, challenge, excitement, frustration, and fun. These are five of the seven emotions that participants rated after each play condition. As such, our AV to emotion model (see Figure 56) had two inputs (arousal and valence), and five outputs (boredom, challenge, excitement, frustration, and fun). Inputs and outputs were represented as percentages of the possible maximum.

8.4.1 Membership Functions

The membership functions and rules for converting arousal and valence into emotion were generated using the Affect Grid, developed from the circumplex model of emotion ([114], see Figure 10). We modified the Affect Grid to have six levels of arousal and valence instead of nine levels (see Figure 10 and Figure 57), Using the modified Affect Grid, we mapped our arousal and valence values from the first model into a language of emotion. We represented arousal and valence in six levels: veryLow, low, midLow, midHigh, high, and veryHigh. As such, our inputs of arousal and valence used six evenly distributed membership functions. Because our mappings from arousal and valence to emotion were based on the six levels, we used trapezoidal membership functions rather than the triangular membership functions employed in the first model. The trapezoidal functions allow for a flat 'roof' on the membership function, rather than a 'point' (see Figure 56). We wanted to remove fuzziness for the input values that were securely in the middle of any given level, and only make use of fuzziness at the boundaries between levels.

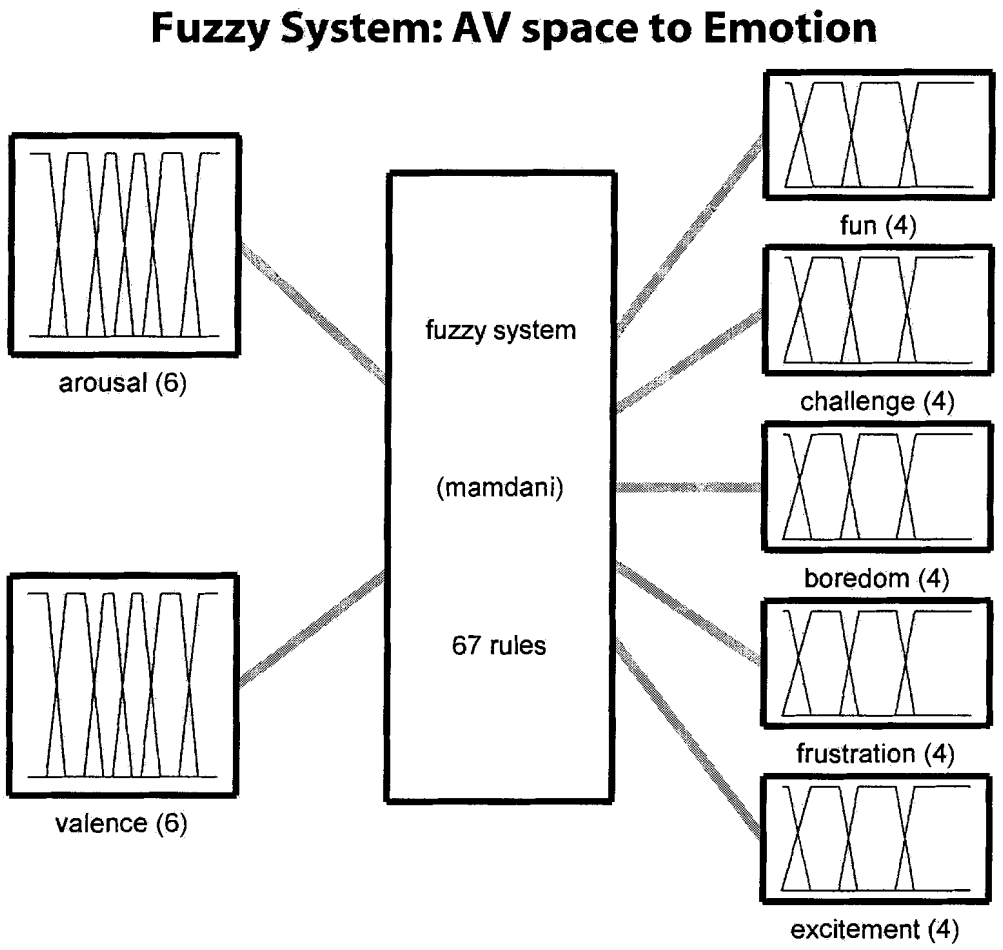


Figure 56: **Modeling emotion from arousal and valence. The number of membership functions applied to that input or output follows the input / output labels. The system used 67 rules to transform the 2 inputs into the 5 outputs.**

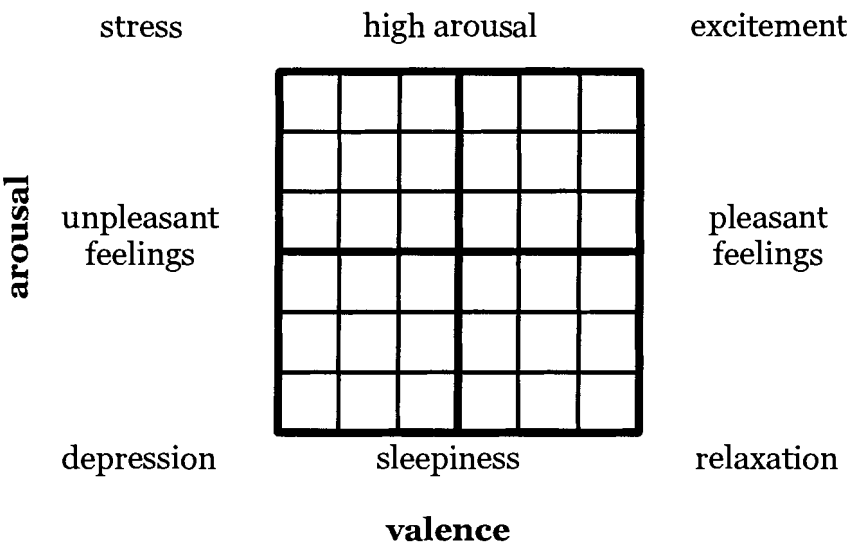


Figure 57: Our interpretation of the Affect Grid: Based on the circumplex model of emotion, the affect grid allows for a quick assessment of mood as a response to stimuli in arousal-valence space [114]. We changed the grid from having nine levels of arousal and valence, to having six levels of arousal and valence.

As shown in Figure 58, we defined the five emotion outputs to have three levels: low, medium, and high, and mapped these levels onto the six levels of AV space. There are no established methods of describing levels of emotions as they vary in AV space. As such, we used guidelines from the labels on the circumplex model of emotion ([114], see Figure 57) to define the levels of fun, challenge, boredom, frustration, and excitement (see Figure 58). The areas in AV space where there was no mapping for a particular emotion was defined as very low for that emotion. As such, our emotion outputs were in four levels: very low, low, medium, and high (Figure 58). As with the inputs, we used trapezoidal membership functions to only make use of fuzziness around the boundaries between levels of modeled emotion (see Figure 56).

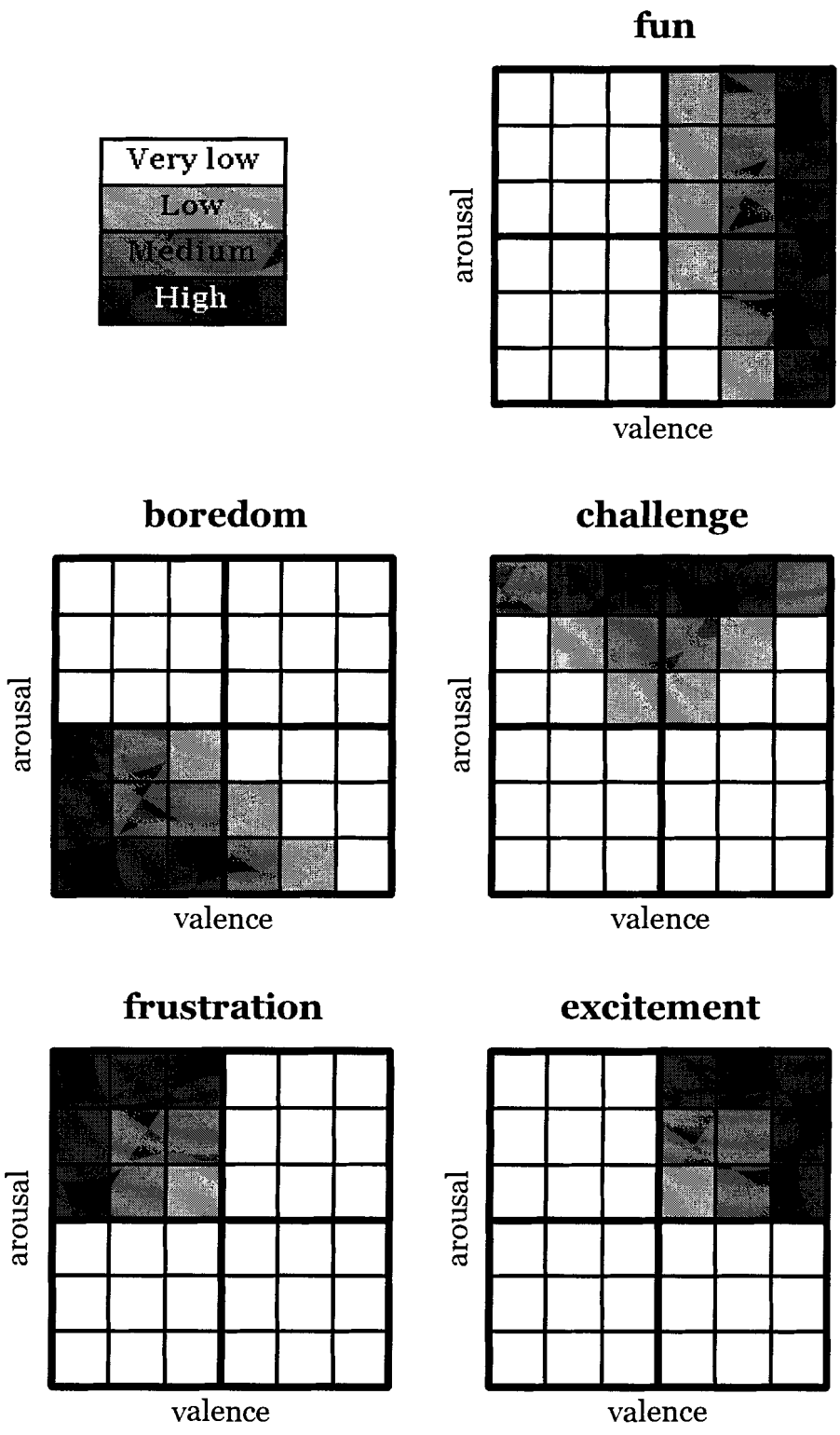


Figure 58: Our representation of levels of emotion in arousal-valence space.

8.4.2 Rules

The rules were generated to simply map the levels of arousal and valence in Figure 58 to the levels of fun, boredom, challenge, frustration, and fun, also shown in Figure 58. Both arousal and valence contributed equally to the generation of, boredom, challenge, excitement, frustration, and fun.

The combination of the membership functions and rules produce the surfaces shown is Figure 59 for the conversion of arousal and valence into fun, boredom, challenge, frustration, and excitement. The 67 rules are presented in Appendix 13.

Because we used data from the six subjects to iteratively generate the model, we will not present the mean results from the emotion model. See Chapter 9 for an analysis of the output of the emotional model for the other six subjects in the experiment.

8.4.3 Issues with Modeling Emotion

The transition from AV space to the five modeled emotions was fairly straightforward. The main issue arises from the fact that there are no established guidelines for transforming levels of arousal and valence to levels of emotion in a continuous manner. We defined the membership functions and the rules to translate AV space to emotion based on the circumplex model of emotion, common sense, and our own understanding of where the five emotions exist in AV space.

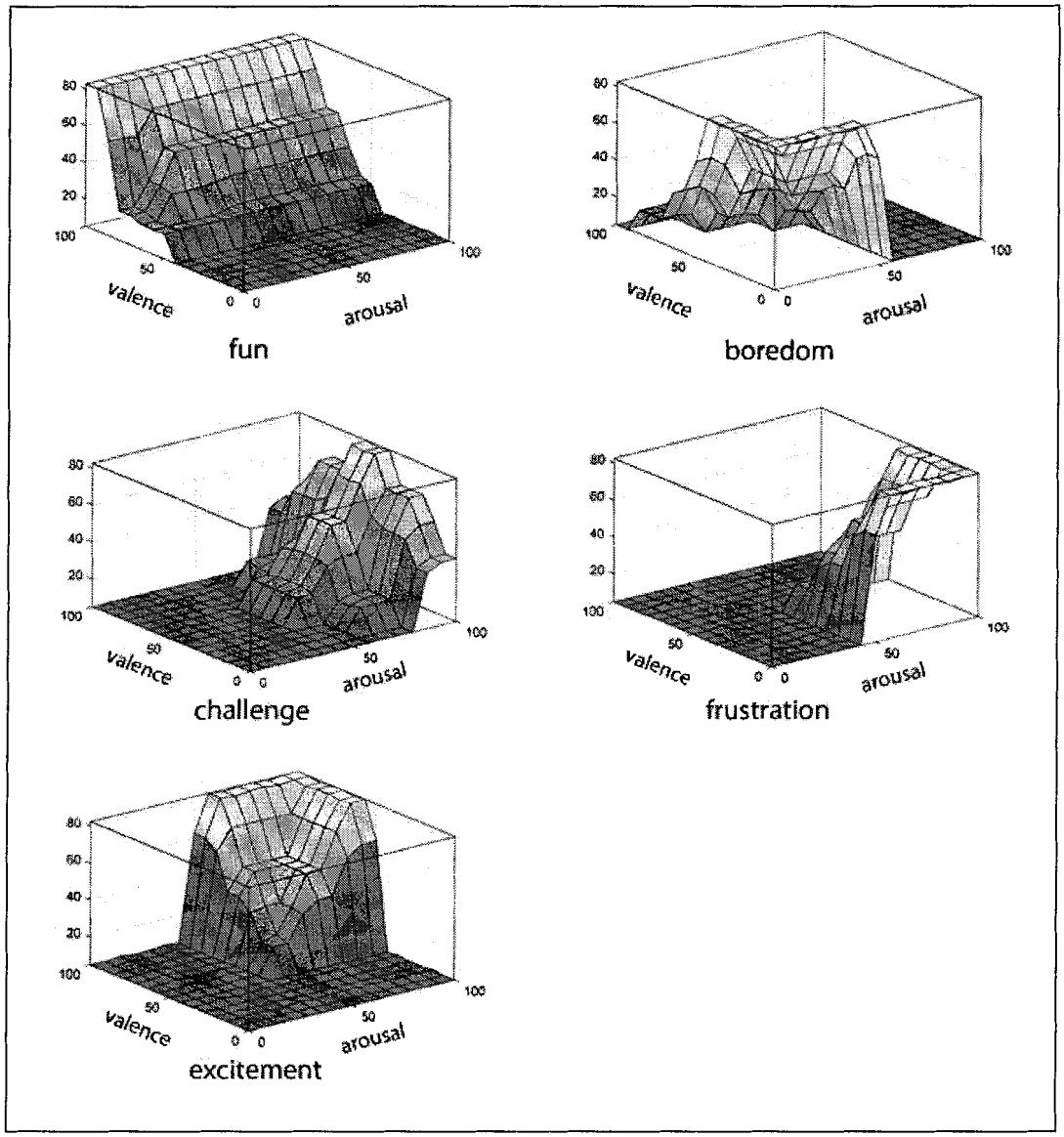


Figure 59: Surfaces showing how arousal and valence are converted into fun, boredom, challenge, frustration, and excitement.

In addition, there are emotions that we wanted to model that aren't easily defined in AV space. We asked subjects to rate their experience along seven subjective dimensions, including ease and engagement. Ease and engagement aren't emotions, and have no well-defined relation to AV space. There are other emotions of interest to

evaluating experience with interactive play technologies such as schadenfreude, naches, and fiero (see section 4.5.1) that aren't easily defined in AV space. How would one use arousal and valence to describe increasing levels of pride in triumphing over adversity or gloating over the misfortune of opponents? More research needs to be conducted to determine how these emotions can be described by arousal and valence before they can successfully be modeled using our fuzzy logic approach.

Chapter 9 USING THE MODEL OF EMOTION

To analyze the effectiveness of our model, we used data gathered from the six subjects not used in the generation of the model. Obtaining successful results using a clean set of data would show the generalizability of our model across individuals, but not across situations or applications.

Data were smoothed and normalized using the previously described method (see Chapter 8.1.4). The physiological signals to AV space and AV space to emotion models were applied to the data and the time series for each emotion were averaged so that we could compare modeled emotion to the subjective responses. Although subjective responses sometimes deviate from actual experience [79, 149], we can use the reported emotions to gauge the accuracy of our model.

9.1 Modeled Emotion

Mean modeled emotions (represented as a percentage) from the six new subjects were analyzed using a repeated measures MANOVA with the five emotions as dependent measures, and play condition as a within-subjects factor. Mean results and statistics are shown in Table 18. Play condition significantly impacted fun and excitement, but not frustration, boredom, or challenge (see Figure 60). Post-hoc analysis revealed that players were having more fun when playing against a friend than when playing against

a stranger (p = .001) or a computer (p = .004), and that playing against a stranger was more fun than playing against a computer (p = .014). Playing against a friend was more exciting than playing against the computer (p=.031), while playing against a stranger was marginally more exciting than playing against the computer (p = .053). There was no difference in excitement between playing against a stranger or a friend (p = .412).

Table 18: Means for modeled emotion, represented as a percentage. There was a significant difference in excitement and fun between play conditions.

	Computer	Friend	Stranger	$F_{2,10}$	p	η2
Boredom	8.5	6.0	6.5	2.7	.118	.35
Challenge	17.3	18.2	22.5	0.55	.594	.10
Excitement	**21.0**	**52.1**	**42.1**	**5.0**	**.032**	**.50**
Frustration	9.7	6.1	7.3	2.4	.145	.32
Fun	**46.7**	**64.2**	**56.9**	**22.1**	**.003**	**.82**

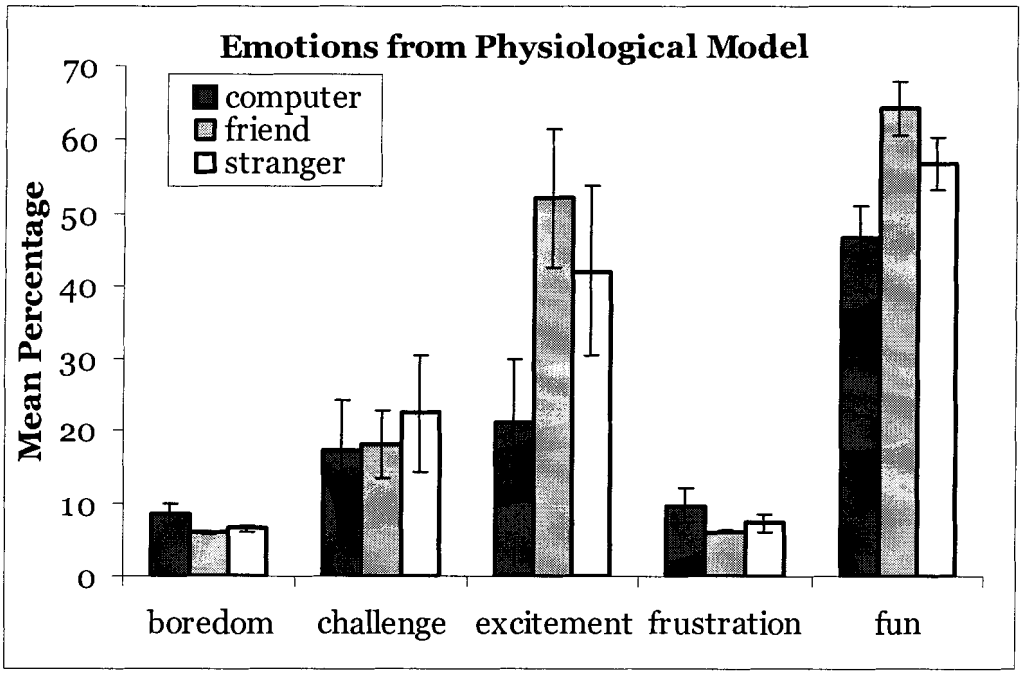


Figure 60: Means (±SE) of modeled emotion, represented as a percentage, separated by play condition.

9.2 Reported Emotion

Participants were asked to rate the boredom, challenge, excitement, frustration, and fun of each condition on a 5-point scale. Mean results and statistics are shown in Table 19. Friedman tests for 3-related samples revealed no differences between conditions (see Figure 61).

Table 19: Means for subjective responses on a 5-point scale. A response of "1" corresponded to "low" and "5" to "high". There were no differences between play conditions.

	Computer	Friend	Stranger	χ^2	Sig.
Boredom	2.2	1.5	2.2	1.4	.504
Challenge	4.2	3.7	3.5	1.6	.444
Excitement	3.7	4.7	4.2	4.5	.104
Frustration	3.5	3.0	2.3	2.5	.291
Fun	4.0	5.0	4.3	5.6	.062

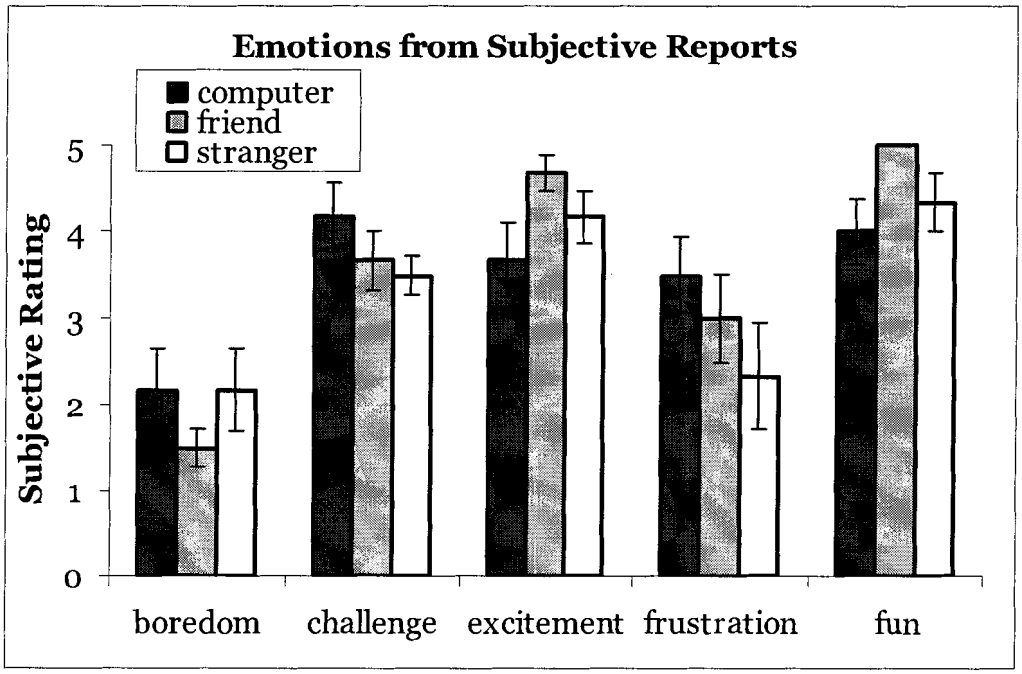


Figure 61: Means (±SE) of the subjective reports on a 5-point scale, separated by play condition.

9.3 Comparing Modeled and Reported Emotion

Although there were no subjective differences between conditions, plotting the means reveals that there were definite trends (see Figure 61). Furthermore, plotting the modeled emotion means reveals the same trends for boredom, excitement, and fun (see Figure 60).

To determine how closely the modeled (objective) emotion resembled reported (subjective) emotion, we correlated the two data sources for each emotional state. We used Spearman's rho, since the subjective reports are non-parametric, while the modeled emotion means are parametric. The subjective and physiological emotional state were significantly correlated for fun ($rho=.99$, $p<.001$), and excitement ($rho=.99$, $p<.001$); the same two emotional states where the model revealed significant differences across play conditions. There was no correlation for boredom ($rho=.50$, $p=.333$) or frustration ($rho=.50$, $p=.333$). Although the same trends were present for reported boredom and modeled boredom, the values for modeled boredom were very low and similar; the same problem existed with frustration. Both of these modeled emotions suffered from issues with scaling, which are discussed later in section 9.4.

There was a correlation for challenge ($rho=.99$, $p<.001$), but the correlation was inverse, as seen in Figure 60 and Figure 61. Subjective ratings for challenge decreased from computer to friend to stranger, while modeled challenge increased from computer to friend to stranger. There were no significant differences in play condition for either modeled or reported challenge; however, the correlation reveals an inverse relationship. In modeling challenge, we assumed that a player's arousal would

increase with challenge; however, upon further examination, this pattern was only true for about half of the participants, while the opposite was true for the other half. Some participants' comments revealed a strategy to attempt to relax when challenged, in order to improve their performance. Obviously, how participants handle challenge in a game is an individual strategy and additional work is required before challenge can be modeled accurately.

We also examined the subjective results from the post-experiment questionnaires. Frequencies of responses for which condition was deemed the most fun, most challenging, and most exciting were tabulated, as were frequencies for the play condition with the maximum modeled fun, challenge, and excitement. For fun, subjective choice and modeled choice were matched for 5 of the 6 (83%) participants; for excitement, subjective choice and modeled choice matched for all 6 (100%) participants. For challenge, only 1 of the 6 (17%) matched. These results corroborate aforementioned mean results for each condition. Participants were not asked which condition they perceived as the most frustrating or boring, thus these emotional states cannot be compared to the post-experiment questionnaires.

9.4 Scaling Issues

Although the trends between conditions are similar for most of the emotions, there are apparent differences in the relative strength of the emotions. Our model represents the emotion as a percentage of the possible maximum and minimum, given the available data. Computer games are generally fun, enjoyable experiences. Although a user may be frustrated, and may rate this frustration as fairly high on a 5-point scale, this

frustration will be low when compared to the frustration experienced by getting a flat tire on the way to an important appointment, or by trying to contact technical support for a lousy local internet provider. By the same logic, the boredom reported by subjects will be much lower than the boredom experienced during a really boring lecture given by a monotonous professor. We asked participants to agree with the statement "this condition was frustrating". Had we asked them to rate their response as a ratio of how frustrating it was compared to a flat tire on the way to an appointment, we probably would have seen much different subjective results. In contrast, our model takes a global approach to the scaling of emotion, so a user's frustration is given as a percentage of the maximum possible frustration, given the available data. As seen in Figure 60 and Figure 61, boredom, challenge, and frustration are significantly lower for modeled emotion than for reported emotion, while fun and excitement are only somewhat lower. This result is expected, since playing a computer game can be quite fun and exciting, but perhaps not as much fun, nor as exciting as riding a rollercoaster or attending a rock concert.

In addition to the scaling issues with subjective reports, sections 8.3.6 and 8.4.3 discuss the scaling issues with the modeled emotions. Although we took a global approach to scaling, given the available data, we cannot be certain that our modeled emotions represent the percentage of the maximum value of each particular emotion exactly. We can only be certain that our values represent percentages of emotion for playing a console game. For example, had we collected GSR, HR, and facial EMG when participants were riding a rollercoaster or dealing with a flat tire, we may have seen different absolute values for our modeled emotions. Using the IAPS to scale

responses in AV space, as discussed in section 8.3.6, may have provided a slightly different scale.

9.5 Modeled Emotion: a Continuous Data Source

Mean modeled emotion is an objective and quantitative metric for evaluating interactive play technologies that reveals variance between conditions. In addition, modeled emotion from physiological data is very powerful as it can continuously and objectively provide a quantitative metric of user experience within a play condition. The mean values shown in Figure 60 are derived from a time series for the five modeled emotions. As such, we can not only see the difference between conditions, but can follow the variance within a condition. Figure 62 shows one participant's modeled frustration over time when playing against a friend and a stranger. The mean values reveal that participant three was most frustrated when playing against the computer, (mean = 19.8%), followed by playing against a stranger (mean=13.1%), and playing against a friend (mean=6.5%). Means alone do not tell us whether the tonic level was raised or whether there were more phasic responses. Modeled emotion pinpoints moments in time when a user's frustration was changing. This is particularly beneficial when there is no baseline or comparative condition. Researchers and developers can uncover individual moments when a user begins to get stressed, starts having fun, or becomes bored.

One of the main drawbacks to using observational analysis is the enormous time commitment associated with watching and annotating hours of video data. Continuously modeling emotion can significantly reduce the amount of time needed to

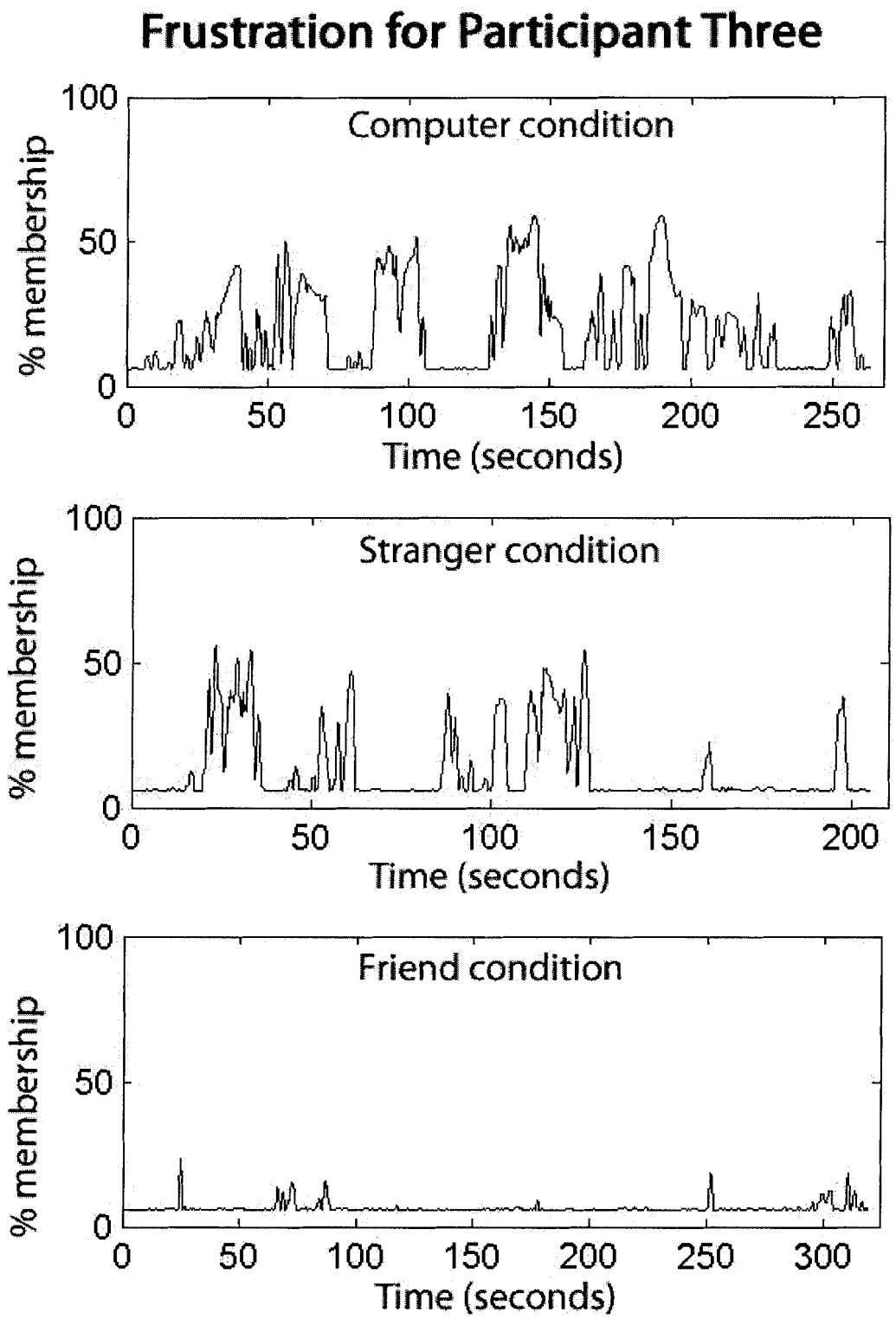


Figure 62: **Frustration for one participant in three conditions. Examining the mean output may reveal differences between conditions; however, examining the entire time series reveals how a participant's emotional state changes over time.**

perform observational analysis. By modeling emotion, researchers can look for interesting features in the emotional experience, then refer to the corresponding video to examine what events preceded the emotional reactions such as increasing boredom, increasing fun, or sustained levels of high frustration.

Researchers could also use continuous emotions to examine how the emotional experiences co-vary. Flow (see section 3.5) refers to an experience state that causes deep enjoyment, due in part to the right balance between the skill of the participant and the challenge of the activity [21]. By monitoring the change in challenge along with corresponding changes in frustration and boredom, researchers can see when players may be in danger of leaving a flow state due to an imbalance between skill and challenge. Continuously modeling emotion can reveal when challenge decreases enough to cause boredom to increase, or conversely, when challenge increases enough to cause frustration to increase. Future research could include using this information to dynamically adjust the challenge of the activity, keeping the player in a state of flow.

9.6 Summary of Modeling Emotion

We used a fuzzy logic approach to transform GSR, HR, $EMG_{smiling}$, and $EMG_{frowning}$ into arousal and valence. The results from the fuzzy model were comparable to a manual approach. In addition, the results were consistent with predictions based on the results from Experiment Two. A second fuzzy model was used to convert arousal and valence into five emotions: fun, challenge, boredom, frustration, and excitement.

Modeled emotion was represented both as an average over a condition, and as a time series that represents an entire condition.

Mean emotion modeled from physiological data provides a metric to fill in the knowledge gap in the objective-quantitative quadrant of evaluating user interaction with entertainment technologies. In addition, the emotion of the user can be viewed over an entire experience, revealing the variance within a condition, not just the variance between conditions. This is especially important for evaluating user experience with entertainment technology, because the success is determined by the *process* of playing, not the *outcome* of playing [96]. The continuous representation of emotion is a powerful evaluative tool that can be easily combined with other evaluative methods, such as video analysis. Given a time series of emotional output, researchers can identify interesting features, such as a sudden increase or decrease in an emotional state, then investigate the corresponding time frame in a video recording. This method would drastically reduce the time required to qualitatively examine video of user interaction with entertainment technologies.

Modeled emotion corresponds to reported emotion for most of the emotions that we investigated. Challenge was an exception that requires additional research on how people differentially respond to challenge in play environments. For the other emotions, the trends were similar between the subjective and objective methods, but the relative strength was different. When modeling emotion, we took the maximum potential experience into consideration, whereas the same was not true of reported emotion. To scale reported emotion, one could choose to ask questions that contained

scaling elements. To better scale modeled emotion, one could collect baseline data using the IAPS.

Chapter 10 SUMMARY AND CONCLUSIONS

10.1 Summary

We have presented a series of experiments to determine the efficacy of using physiological signals as indicators of emotional experience with entertainment technologies. Chapter 2 demonstrated that there is a lack of objective and quantitative evaluation methodologies for studying user interaction with play technologies. The three presented experiments advance the understanding of body reactions to play technologies, and move towards an objective and quantitative methodology of evaluation.

Experiment One was an exploration of how physiological signals respond to interaction with play technologies. Experiment Two investigated how physiological signals co-vary with subjective reports. Based on the knowledge acquired in Experiments One and Two, Experiment Three presented a method for modeling emotion, using physiological signals. The modeled emotions were successfully compared to subjective reports. The summaries and contributions of each phase of the presented research follow.

10.1.1 Experiment One: Goldilocks

Our first experiment was designed to explore how physiology can be used to objectively measure user experience with entertainment technology. Prior to the first experiment, we only had theoretical information, based on the literature, on how the body would respond to play environments. We collected a variety of physiological measures (GSR, EKG, EMG_{jaw}, respiration) while observing participants playing NHL 2003. Participants played in four difficulty conditions (beginner, easy, medium, and difficult), to either create an experience that was too easy, that was too hard, or that matched a player's experience to the difficulty level in the game, creating a condition that was 'just right'. We expected that participants would prefer playing in the condition that was best matched to their level of expertise, and that these preferences would be reflected in their subjective experience as well as their physiological experience.

The chosen experimental manipulation did not produce consistent subjective results across all participants. We saw no differences in boredom, frustration, or fun across difficulty conditions, and only reported challenge increased with increases in difficulty. Without consistent subjective results, we did not expect consistent physiological results, and determined that our experimental manipulation was not appropriate for exploration of how the body responds to interactive play environments. In addition, further analyses uncovered some methodological issues that contributed to irregular patterns of physiological activity. Primarily, the act of conducting the experiment produced different phases in the experiment (e.g., play, interview, rest) that created greater physiological responses than the experimental

manipulations themselves. In addition, we could not remove artifacts from the upwards drift in GSR over time.

Although Experiment One did not produce interesting subjective or physiological results, we were able to achieve our goal of exploring how the body responds to interactive play environments. We were also able to generate some rules for conducting experiments in this domain that aided us in our subsequent experiments. Having corrected the methodological issues, we designed a second experiment with a different experimental manipulation that we felt would produce a consistent experience for all players.

10.1.2 Experiment Two: Turing

We conducted a second study to further understand how body responses can be used to create an objective evaluation methodology. Because this methodology is a novel approach to evaluate play technologies, and the results from Experiment One were ambiguous, we used an experimental manipulation designed to maximize the difference in the experience for the participant. The participants played in two conditions: against a co-located friend, and against the computer.

We chose these play conditions because we have previously observed pairs (and groups) of participants playing together under a variety of collaborative conditions [22, 54, 75, 120]. Our previous observations revealed that players seem to be more engaged with a game when another co-located player is involved. Thus, we thought that participants would be more excited, have more fun, and prefer playing against a

friend than when playing against a computer. Additionally, we hypothesized that differences in the participants' subjective experiences would be reflected in their physiological activities.

After addressing our methodological issues from Experiment One, Experiment Two tested and supported four experimental hypotheses. We found that participants preferred playing against a friend to playing against a computer; participants experienced higher GSR values when playing against a friend than against a computer; participants experienced higher EMG values along the jaw when playing against a friend than against a computer; and the differences in the participants' GSR signal in the two conditions was correlated to the differences in their subjective responses for fun. We also found other correlations between the normalized subjective measures and the normalized physiological measures. The confirmation of our hypotheses provided support for our two main conjectures: that physiological measures can be used as objective indicators for the evaluation of co-located, collaborative play; and that the normalized physiological results will correspond to subjective reported experience.

Experiment Two showed that when a physiological time series is averaged for each condition, mean values yield meaningful results that respond in a similar manner to subjective reports. These results have the same disadvantage as subjective results, in that they are single points of data representing an entire condition; however, unlike subjective reporting, they represent an *objective* measure of user experience. Used in

concert with subjective reporting, the two methods can provide a more detailed and accurate representation of the player's experience.

The raised average GSR when playing against a friend revealed that players were more aroused when playing against a friend than when playing against a computer. However, Experiment Two did not show whether this elevated result is due to a higher tonic level or more phasic responses. Physiological data provide a high-resolution time series that can discriminate between experiences with greater resolution than averages alone. In Experiment Two, we graphically represented continuous responses to different game events. In the next experiment, we wanted to take advantage of the high-resolution, contextual nature of physiological data to provide an objective, and continuous measure of player experience.

10.1.3 Experiment Three: Modeling Emotion

Experiment Three was designed to test the conjecture that physiological metrics could be used to model user emotional experience when playing a game, providing continuous, quantitative, and objective metrics of evaluation for interactive play technologies. Therefore, our third experiment presented a method of modeling user emotional state when interacting with play technologies. Due to the success of Experiment Two in separating responses from playing against a computer versus playing against a friend, we continued this approach and added a third condition (playing against a stranger. Thus, we collected data in three play conditions: against a co-located friend, against a co-located stranger, and against the computer. Using the entire time series, we developed a fuzzy logic model that transformed four

physiological signals (GSR, HR, $EMG_{smiling}$, and $EMG_{frowning}$) into values of arousal and valence. The output from the model conformed to expected values for each play condition. In addition, modeled arousal and valence were similar, but superior to a brute force approach of calculating arousal and valence. A second fuzzy logic model transformed the arousal and valence values into continuous values for five emotions: boredom, challenge, excitement, frustration, and fun.

The modeled emotions show the same trends as reported emotions for fun, boredom, and excitement; however, the modeled emotions revealed differences between play conditions, while the differences between the subjective reports failed to reach significance. Modeled challenge did not correspond to reported challenge, and more research is needed to understand how people physiologically respond to challenging play environments.

Mean emotion modeled from physiological data provides a metric to fill in the knowledge gap in the objective-quantitative quadrant of evaluating user interaction with entertainment technologies. Figure 63 shows that there are several choices in methodologies for evaluating user interaction with play technologies, but that there are no appropriate techniques for objective and quantitative evaluation since task performance metrics aren't relevant to play. Heuristic evaluation could be seen as a quantitative methodology since experts can provide ratings for how well software adheres to the heuristics. Observational analysis is a tool that can be used to generate quantitative or qualitative results, but is not used quantitatively to evaluate entertainment technologies due to the time commitment and expertise needed. Figure

64 shows how modeled emotions provide an alternative evaluation methodology for researchers interested in a quantitative and objective evaluation.

In addition, the emotion of the user can be viewed over an entire experience, revealing the variance within a condition, not just the variance between conditions. This is especially important for evaluating user experience with entertainment technology, because success is determined by the *process* of playing, not the *outcome* of playing [96].

10.2 Thesis Contributions

The goal of this research is to investigate the efficacy of physiological signals as indicators of user experience with interactive play technologies. In the course of this work, we have made significant contributions to affective computing, HCI evaluation methodologies, and extended the applicability of fuzzy logic to a new domain. Specific contributions are outlined in this section.

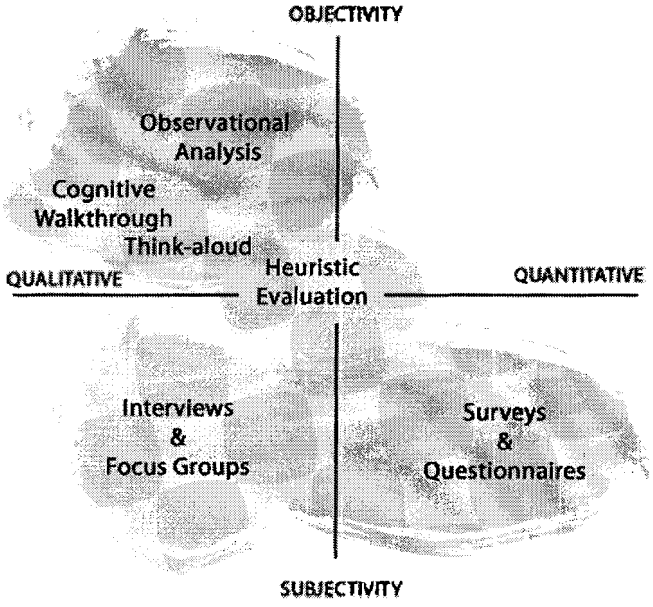


Figure 63: Current methods for evaluating entertainment technologies. Evaluators have a lot of choice, but there is a knowledge gap in the quantitative-objective quadrant since task performance metrics aren't relevant.

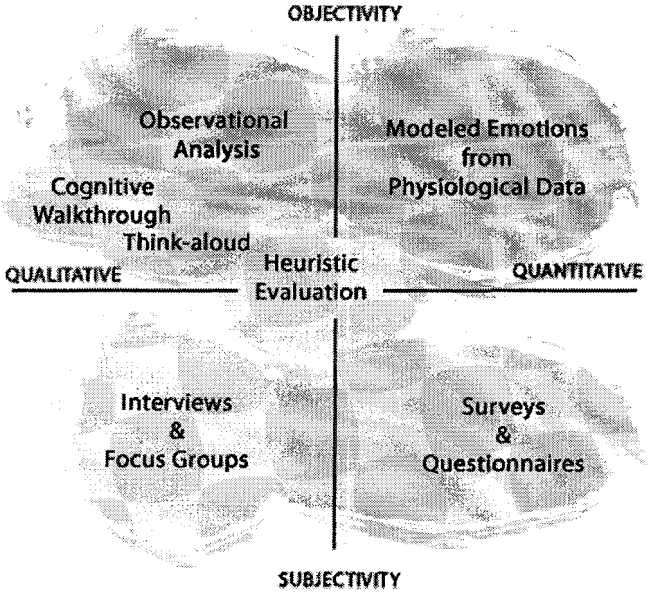


Figure 64: Contribution of this dissertation. There was a knowledge gap in the quantitative-objective quadrant, since task performance metrics were not used. Modeled emotions from physiological data fill this quadrant, providing a new choice for evaluators of entertainment technologies.

10.2.1 Systematic Exploration of How the Body Responds to Interactive Play Environments

In Experiment One, we systematically examined how a user's physiological signals respond to changes in difficulty level when interacting with a computer game. Although physiological signals have been used extensively as indicators of mental effort and stress, there has been no previous research investigating how the body responds to interactive play technologies.

10.2.2 Rules and Guidelines for Conducting Research in this Domain

Although our participants did not respond consistently to the changes in difficulty level, the first experiment revealed issues in our methodology that were potentially confounding our results. When examining play environments, researchers have to deal with unique issues, not apparent when examining typical productivity software. For example, variability of game intensity is incorporated into game design as a method of pacing the play experience. Collapsing a time series into a single point erases the variance within each condition, causing researchers to lose valuable information. In addition, participants create ways to enjoy themselves in all experimental conditions, which should impact the researcher's choice of experimental manipulation.

Physiological metrics have high individual variability, making comparison across subjects impossible without some form of normalization. Also, physiological metrics are highly responsive signals, so resting periods must be incorporated into the

experiment design in order to allow the signals to return to baseline levels prior to each experimental condition. Even with resting periods, order effects can remain and researchers need to acknowledge effects due to order. Finally, the act of applying sensors to the body and monitoring body responses can be a stressful experience for a participant, and every effort must be made to allow the participant to relax and feel at ease.

10.2.3 Physiological Measures Can be Used to Objectively Measure a Player's Experience with Entertainment Technology

In Experiment Two, we found evidence that there is a different physiological response when playing against a computer versus playing against a friend. Examining the means of the physiological signals revealed elevated levels of GSR and EMG of the jaw when playing against a friend. The mean results do not tell us whether the tonic level of the signal is elevated or whether there are more phasic responses. Physiological data provide a high-resolution time series that can discriminate between experiences with greater resolution than averages alone. In Experiment Two, we graphically represented continuous responses to goals and fights in the game. By windowing and graphing GSR, we saw bigger reactions to goals and fights when playing against a friend over playing against a computer. This windowing technique was the first step towards making use of the continuous nature of physiological signals.

10.2.4 Normalized Physiological Measures of Experience with Entertainment Technology Correspond to Subjective Reports

In Experiment Two, we found many correlations between normalized physiological measures and normalized subjective reports. Normalizing and correlating the data, as we did in Experiment Two, is a powerful tool because it shows that the *amount* by which participants increased their subjective ratings corresponded to the *amount* by which their mean physiological data increased. In addition, this approach contains results that may otherwise get lost. ANOVAs show results when all participants are responding in a similar manner, however correlations will reveal patterns even when participants are responding differently from one another, a useful tool when investigating something as individual as engagement with play technologies.

10.2.5 A Method of Modeling Emotion

The results that we gathered in the first two experiments formed a basis for developing a model of user emotion, based on physiological reactions. Our first model transformed physiological signals into AV space. Representing a participant's experience in AV space is a great method of objectively and quantitatively measuring their experience when engaged with entertainment technologies. We graphed the participant's experience continuously in AV space to determine how our model compared to a manual approach. These graphs visually represent the positive and negative stimulation that the participant feels as they engage with the technology.

We moved beyond AV space by creating a method to transform AV space into five emotions: boredom, challenge, excitement, frustration, and fun. Emotions modeled from physiological data provide a metric to fill the knowledge gap in the objective-quantitative quadrant of evaluating user interaction with entertainment technologies. The modeled emotions were compared to subjective reports and showed the same trends for fun, boredom, and excitement; however, modeled emotions revealed differences between play conditions, while the differences between the subjective reports failed to reach significance.

Our modeled emotions were based on fuzzy transformation functions from physiological variables to AV space and then from AV space to emotions. We based our decisions for the membership functions and rules on a theoretical understanding of how the physiological signals operate, and how we expect users to feel when playing a game. Although other mappings could be considered, our results provide a proof of concept of the modeling technique. In addition, integrating data from more participants engaged in a broader range of play situations could improve our mappings from physiological metrics to emotion.

10.2.6 Modeled Emotions Provide a Continuous Metric for Evaluation

In addition to providing an objective and quantitative approach to evaluating play technologies, modeled emotion can be viewed over an entire experience, revealing the variance within a condition, not just the variance between conditions. This is especially important for evaluating user experience with entertainment technology,

because the success during play is determined by the *process* of playing, not the *outcome* of playing [96]. Continuously representing emotion is a powerful evaluative tool that can be easily combined with other methods. Given a time series of emotional output, researchers can use interesting features in the modeled emotion output to index other evaluative data sources such as video or screen captures of the play environment.

10.3 Future Work

Although our modeled emotions correspond well to reported emotion, there are still improvements to the model that could be made. First, the scaling of the arousal and valence axes could be improved. In order to determine maximum arousal and valence, we used the minimum and maximum values from the all three play conditions and the rest period. We determined the baseline arousal and valence values to the best of our ability, given the available data; however, the available data may not have contained accurate baseline values. A better approach to scaling the arousal and valence axes would have been to use the IAPS [64] to baseline participants' arousal and valence by presenting pictures from the IAPS data set, and measuring a subject's physiological responses. Since GSR is not consistent across experimental sessions, baselining a participant's GSR response on one day might not apply to the following day or week. Using a variety of baselines and dynamically adjusting for the day-to-day variations would be a feasible approach, but would require additional research in order to be implemented correctly.

We developed models for five emotional states that we felt were relevant to interaction with entertainment technology. We would like to consider other relevant emotional states that can be described by arousal and valence, such as disappointment, anger, or pride. Other emotions, such as schadenfreude or fiero might be more difficult to describe in terms of arousal and valence, and more research needs to be conducted on these emotions which are less easily defined.

Of the five emotional states that we modeled, boredom, excitement and fun compared well to reported emotions through subjective responses. Our model of challenge produced results that were in direct opposition to reported challenge. We thought that challenge could be modeled mainly through increasing arousal and neutral valence; however, it was made known from the comments that some participants responded to increasing challenge by actively trying to relax in order to improve their performance. Further work needs to be conducted on how people respond to challenge and frustration in play environments for these emotions to be effectively modeled.

In addition to comparing the modeled emotions to subjective reports, we would like to relate them to another objective data source, gathered through observational analysis. Facial expressions, verbalizations, or game events could be used to connect the emotional responses to events that were occurring in the context of play.

Along these lines, we would like to investigate how we can combine modeled emotions with other evaluation methods to produce a better, more complete understanding of a player's interaction with play technologies. For example, we could use modeled emotions to reduce the time commitment associated with observational

video analysis. Given a time series of emotional output, researchers can identify interesting features, such as a sudden increase or decrease in an emotional state, then investigate the corresponding time frame in a video recording. This method would drastically reduce the time required to qualitatively examine video of user interaction with entertainment technologies. Modeled emotions could also be used in conjunction with other methods of evaluation, such as heuristics.

In addition, we would like to see if our method can generalize to interaction with other play technologies, specifically, to study user behaviour in ubiquitous play [9, 71] environments. In our earlier studies [120], described in section 2.3.4, we used more traditional methods of evaluating user interactions with the technology and with other players. These methods, including subjective reports and observational analysis fell short due in part to limited evaluative bandwidth. When determining how to evaluate play environments that used emerging technologies, such as the False Prophets game environment [76], there was no comparative environment that could compete in terms of novelty. Traditional methods were not robust enough to evaluate our novel ubiquitous play environments. We regard modeled emotions as a means to successfully evaluate novel play environments objectively, quantitatively, and continuously. As such, we plan to conduct more research on applying the methods for modeling emotion to the evaluation of ubiquitous play environments. This includes the introduction of mobility, the use of less invasive sensors, and an algorithmic approach to contend with the effects of physical activity.

We have demonstrated how modeled emotions can be used as an evaluative tool, but they could also be used to dynamically adapt play environments to keep users engaged. Flow (see section 3.5) refers to an experience state that causes deep enjoyment, due in part to the right balance between the skill of the participant and the challenge of the activity [21]. By monitoring the change in challenge along with corresponding changes in frustration and boredom, researchers could see when players were in danger of leaving a flow state due to an imbalance between skill and challenge.

Finally, the techniques described in this paper could be adapted to analyze a user's emotional response to productivity software, or other work-related interactive technologies. Although task performance is used to objectively and quantitatively assess interaction with productivity technologies, modeled emotions have a high evaluative bandwidth, not seen in many other evaluation methodologies.

10.4 Conclusions

Researchers are using emerging technologies to develop novel play environments, while established computer and console game markets continue to grow rapidly. Even so, we have demonstrated how evaluating the success of interactive play environments is still an open research challenge. Traditional evaluation methods have been adopted, with some success, for quantitative-subjective, qualitative-subjective, and qualitative-objective assessment of play technologies. While performance metrics are used for quantitative-objective analysis of productivity systems, the success of play environments is determined by the *experience of playing*, not the *performance of the*

participant. As such, there is a knowledge gap for quantitative-objective evaluation of play technologies. In addition, the existing techniques suffer from low evaluative bandwidth.

We have presented a series of three experiments, based on physiological signals, that generate a model of user emotion for interaction with play technologies. Modeled emotions can be a powerful evaluation technique because they:

1. capture usability and playability through metrics relevant to ludic experience;
2. account for user emotion;
3. are quantitative and objective; and
4. can be represented continuously.

In Experiment One, we explored how a user's physiological signals respond to interaction with play technologies. The results allowed us to generate rules for conducting experiments in this domain. In Experiment Two, we investigated whether physiological signals could differentiate between play conditions, and how physiological signals co-vary with subjective reports. We found evidence that there is a different physiological response in the body when playing a computer game against a co-located friend versus playing against a computer. When normalized, many physiological results were mirrored in the subjective reports. Our results provided support for Conjecture A, that *physiological measures can be used to objectively measure a player's experience with entertainment technology,* and Conjecture B, that *normalized physiological measures of experience with entertainment technology will correspond to subjective reports.*

In Experiment Three, we presented a method for modeling emotion using physiological data. We developed a fuzzy logic model that transformed four physiological signals into arousal and valence. The output from the model conformed to expected values for each play condition (against a computer, a co-located friend, or a co-located stranger). A second fuzzy logic model transformed arousal and valence into five emotions: boredom, challenge, excitement, frustration, and fun. When evaluated with a test data set, our modeled emotions showed the same trends as reported emotions for fun, boredom, and excitement; however, modeled emotions revealed differences between three play conditions, while the differences between reported emotions failed to reach significance. These results support Conjecture C, that *physiological metrics can be used to model user emotional experience when playing a game, providing continuous, quantitative, and objective metrics of evaluation for interactive play technologies.*

Mean emotion modeled from physiological data fills a knowledge gap for objective and quantitative evaluation of user interaction with entertainment technologies. In addition, user emotion can be viewed continuously over an entire experience, revealing variance within a condition, not just variance between conditions. This is especially important for evaluating play experiences, because success is determined by the *process* of playing, not the *outcome* of playing. The continuous representation of modeled emotion is a powerful evaluative tool that can be combined with other approaches for a robust method of evaluating user interaction with play technologies.

Chapter 11 REFERENCES

[1] *Affective jewelry*. Retrieved January 2004, from http://affect.media.mit.edu/AC_research/projects/affective_jewelry.html

[2] *Affquake*. Retrieved January 2004, from http://affect.media.mit.edu/AC_research/sensing.html

[3] ALLANSON, J. (2002). Electrophysiologically interactive computer systems. *IEEE Computer, 35*(3), 60-65.

[4] BAKER, K., GREENBERG, S., & GUTWIN, C. (2001). Heuristic evaluation of groupware based on the mechanics of collaboration. In M. Little & L. Nigay (Eds.), *Engineering for human -computer interaction* (Vol. Lecture Notes in Computer Science 2254, pp. 123-139): Springer-Verlag.

[5] BAKER, K., GREENBERG, S., & GUTWIN, C. (2002). *Empirical development of a heuristic evaluation methodology for shared workspace groupware.* Paper presented at Conference on Computer Supported Cooperative Work, New Orleans, LA, USA, pp. 96-105.

[6] BENFORD, S., ROWLAND, D., FLINTHAM, M., DROZD, A., HULL, R., REID, J., MORRISON, J., & FACER, K. (2005). *Life on the edge: Supporting collaboration in location-based experiences.* Paper presented at Conference on Human Factors in Computing Systems, Portland, Oregon, USA, pp. 721-730.

[7] BERSAK, D., MCDARBY, G., AUGENBLICK, N., MCDARBY, P., MCDONNELL, D., MCDONALD, B., & KARKUN, R. (2001). *Intelligent biofeedback using an immersive competitive environment.* Paper presented at UBICOMP 2001 Workshop on Ubiquitous Gaming.

[8] BJÖRK, S., FALK, J., HANSSON, R., & LJUNGSTRAND, P. (2001). *Pirates! Using the physical world as a game board.* Paper presented at Interact 2001, Tokyo, Japan.

[9] BJÖRK, S., HOLOPAINEN, J., LJUNGSTRAND, P., & MANDRYK, R. L. (2002). Introduction to special issue on ubiquitous games. *Personal and Ubiquitous Computing, 6*, 358–361.

[10] BOUCSEIN, W. (1992). *Electrodermal activity*. New York: Plenum Press.

[11] BOUTCHER, S. H., NUGENT, F. W., MCCLAREN, P. F., & WELTMAN, A. L. (1998). Heart period variability of trained and untrained men at rest and during mental challenge. *Psychophysiology, 35*, 16-22.

[12] CACIOPPO, J. T., BERNTSON, G. G., LARSEN, J. T., POEHLMANN, K. M., & ITO, T. A. (2000). The psychophysiology of emotion. In J. M. Haviland-Jones (Ed.), *Handbook of emotions*. New York: The Guilford Press.

[13] CACIOPPO, J. T., & TASSINARY, L. G. (1990). Inferring psychological significance from physiological signals. *American Psychologist, 45*(1), 16-28.

[14] CACIOPPO, J. T., & TASSINARY, L. G. (1990). Psychophysiology and psychological inference. In L. G. Tassinary (Ed.), *Principles of psychophysiology: Physical, social, and inferential elements*. Cambridge: Cambridge University Press.

[15] CARD, S. K., MORAN, T. P., & NEWELL, A. (1980). The keystroke-level model for user performance with interactive systems. *Communications of the ACM, 23,* 396-410.

[16] CHEN, D., & VERTEGAAL, R. (2004). *Using mental load for managing interruptions in physiologically attentive user interfaces.* Short Paper presented at Conference on Human Factors in Computing Systems, Vienna, Austria, pp. 1513-1516.

[17] CNOSSEN, F., ROTHENGATTER, T., & MEIJMAN, T. (2000). Strategic changes in task performance in simulated car driving as an adaptive response to task demands. *Transportation Research, F*(3), 123-140.

[18] CORNETT, S. (2004). *The usability of massively multiplayer online roleplaying games: Designing for new users.* Paper presented at Conference on Human factors in computing systems, Vienna, Austria, pp. 703-710.

[19] COSTANZA, E., INVERSO, S. A., & ALLEN, R. (2005). *Toward subtle intimate interfaces for mobile devices using an EMG controller.* Paper presented at Conference on Human Factors in Computing Systems, Portland, Oregon, USA, pp. 481-489.

[20] COX, E. (1992). Fuzzy fundamentals. *IEEE Spectrum, 29*(10), 58-61.

[21] CSIKSZENTMIHALYI, M. (1990). *Flow: The psychology of optimal experience.* New York: Harper Perennial.

[22] DANESH, A., INKPEN, K. M., LAU, F., SHU, K., & BOOTH, K. S. (2001). *Geney: Designing a collaborative activity for the palm handheld computer.* Paper presented at Conference on Human Factors in Computing Systems, Seattle, WA, USA, pp. 388-395.

[23] DE WAARD, D. (1996). *The measurement of drivers' mental workload.* The Traffic Research Center VSC: University of Groningen, Haren, The Netherlands.

[24] DESURVIRE, H., CAPLAN, M., & TOTH, J. A. (2004). *Using heuristics to evaluate the playability of games.* Short Paper presented at Conference on Human Factors in Computing Systems, Vienna, Austria, pp. 1509-1512.

[25] DI MICCO, J. M., LAKSHMIPATHY, V., & FIORE, A. T. (2002). *Conductive chat: Instant messaging with a skin conductivity channel.* Short Paper presented at Computer Supported Cooperative Work, New Orleans, LA, USA., pp. 193-194.

[26] *Dictionary.Com.* Retrieved January 2004, from www.dictionary.com

[27] DIENSTBIER, R. A. (1984). The role of emotion in moral socialization. In R. Zajonc (Ed.), *Emotions, cognition, and behavior* (pp. 484-514). Cambridge: Cambridge University Press.

[28] DODD, J., & ROLE, L. W. (1997). The autonomic nervous system. In E. R. Kandel, J. H. Schwartz & T. M. Jessel (Eds.), *Principles of neural science* (3rd ed., pp. 761-791). Norwalk, Connecticut: Appleton & Lange.

[29] EKMAN, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion.* Sussex: John Wiley & Sons, Ltd.

[30] EKMAN, P., & FRIESEN, W. V. (1984). *Emotion facial action coding system (EM-FACS).* San Francisco: University of California.

[31] EKMAN, P., LEVENSON, R. W., & FRIESEN, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science, 221*(4616), 1208-1210.

[32] *Elspa.* Retrieved August 2002, from http://www.elspa.com/

[33] ERICSSON, K. A., & SIMON, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge: The MIT Press.

[34] FEDEROFF, M. A. (2002). *Heuristics and usability guidelines for the creation and evaluation of fun in video games.* Unpublished Master of Science, Department of Telecommunications: Indiana University, Bloomington, IN.

[35] FISHER, C., & SANDERSON, P. (1996, March 1996). Exploratory data analysis: Exploring continuous observational data. *Interactions, 3,* 25-34.

[36] FITTS, P. M. (1954). The information capacity of the human motor system in controlling amplitude of movement. *Journal of Experimental Psychology, 47*, 381-391.

[37] FITTS, P. M., & PETERSON, J. R. (1964). Information capacity of discrete motor responses. *Journal of Experimental Psychology, 67*, 103-112.

[38] FOURNIER, L. R., WILSON, G. F., & SWAIN, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: Manipulations of task difficulty and training. *International Journal of Psychophysiology, 31*(2), 129-145.

[39] FRIDLUND, A. J., & CACIOPPO, J. T. (1986). Guidelines for human electromyographic research. *Psychophysiology, 23*, 567-589.

[40] FROHLICH, D., P., D., & MONK, A. (1994). Management of repair in human-computer interaction. *Human-Computer Interaction, 9*(3), 385-425.

[41] FULTON, B., & MEDLOCK, M. (2003). *Beyond focus groups: Getting more useful feedback from consumers.* Paper presented at Game Developer's Conference, San Jose, CA.

[42] *Galvactivator.* Retrieved January 2004, from http://www.media.mit.edu/galvactivator/

[43] GAMBOA, M., KOWALEWSKI, R., & ROY, P. (2004). *Playtesting strategies.* Paper presented at Game Developer's Conference, San Jose, CA.

[44] *Geogia tech smart shirt.* Retrieved January 2004, from http://www.gatech.edu/news-room/archive/news_releases/sensatex.html

[45] GUTWIN, C., & GREENBERG, S. (2000). *The mechanics of collaboration: Developing low cost usability evaluation methods for shared workspaces.* Paper presented at WETICE 2000, NIST,Gaithersburg, MD USA.

[46] HEALEY, J., DABEK, F., & PICARD, R. W. (1998). *A new affect-perceiving interface and its application to personalized music selection.* Paper presented at Workshop on Perceptual User Interfaces, San Francisco, CA, USA.

[47] HEALEY, J., & PICARD, R. (2000). *Smart car: Detecting driver stress.* Paper presented at The 15th International Conference on Pattern Recognition, Barcelona, Spain.

[48] HEALEY, J., SEGER, J., & PICARD, R. W. (1999). *Quantifying driver stress: Developing a system for collecting and processing bio-metric signals in natural*

situations (tr#483). Paper presented at The Rocky Mountain Bio-Engineering Symposium.

[49] HERZ, J. C. (2002). The bandwidth capital of the world. *Wired Magazine, 10 (8).*

[50] HJELM, S. I. (2003). The making of brainball. *Interactions, 10,* 26-34.

[51] HOLMQUIST, L. E., FALK, J., & WIGSTRÖM, J. (1999). Supporting group collaboration with inter-personal awareness devices. *Journal of Personal Technologies, 3*(1-2).

[52] HOLTZBLATT, K., & JONES, S. (1993). Conducting and analyzing a contextual interview. In D. Schuler & A. Namioka (Eds.), *Participatory design: Principles and practices* (pp. 177–210). London: Lawrence Erlbaum.

[53] HUTT, C. (1979). Exploration and play. In B. Sutton-Smith (Ed.), *Play and learning* (pp. 175-194). New York: Gardner Press.

[54] INKPEN, K., BOOTH, K. S., KLAWE, M., & UPITIS, R. (1995). *Playing together beats playing apart, especially for girls.* Paper presented at Computer Supported Collaborative Learning.

[55] *Interactive digital software association.* Retrieved August, 2005, from www.idsa.com

[56] JORNA, P. G. A. M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology, 34,* 237-258.

[57] KAEHLER, S. D. (March, 1998). *Fuzzy logic tutorial.* In Encoder: Newsletter of the Seattle Robotics Society Retrieved September, 2005, from http://www.seattlerobotics.org/encoder/mar98/fuz/flindex.html

[58] KALSBEEK, J. W. H., & ETTEMA, J. H. (1963). Scored regularity of the heart rate pattern and the measurement of perceptual or mental load. *Ergonomics, 6*(306).

[59] KALSBEEK, J. W. H., & SYKES, R. N. (1967). Objective measurement of mental load. *Acta Psychologica, 27,* 253-261.

[60] KELLY, J. P., & DODD, J. (1997). Anatomical organization of the nervous system. In E. R. Kandel, J. H. Schwartz & T. M. Jessel (Eds.), *Principles of neural science* (3rd ed., pp. 270-282). Norwalk, Connecticut: Appleton & Lange.

[61] KELTNER, D., & EKMAN, P. (2000). Facial expression of emotion. In J. M. Haviland-Jones (Ed.), *Handbook of emotions* (2nd ed.). New York: The Guilford Press.

[62] KWAN, G. (2002). Play attention! *Berkeley Medical Journal Online.*

[63] LANG, P. J. (1995). The emotion probe. *American Psychologist, 50*(5), 372-385.

[64] LANG, P. J., GREENWALD, M. K., BRADLEY, M. M., & HAMM, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioural reactions. *Psychophysiology, 30*, 261-273.

[65] LAZZARO, N. (2004). *Why we play games: 4 keys to more emotion.* Paper presented at Game Developer's Conference.

[66] *Lego mindstorms community bulletin boards.* Retrieved January 2004, from http://mindstorms.lego.com/eng/forums/

[67] LEVENSON, R. W. (1992). Autonomic nervous system differences among emotions. *American Psychological Society, 3*(1), 23-27.

[68] LEWIS, C., POLSON, P., WHARTON, C., & RIEMAN, J. (1990). *Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces.* Paper presented at Conference on Human Factors in Computing Systems, Seattle, WA, USA., pp. 235-242.

[69] *Lifeshirt system.* Retrieved January 2004, from http://www.vivometrics.com/

[70] MACK, R. L., & NIELSEN, J. (1994). Executive summary. In J. Nielsen (Ed.), *Usability inspection methods* (pp. 1-23). New York: John Wiley and Sons, Inc.

[71] MAGERKURTH, C., CHEOK, A. D., MANDRYK, R. L., & NILSEN, T. (2005). Pervasive games: Bringing computer entertainment back to the real world. *ACM Computers in Entertainment, 3*(3), Article 4A.

[72] MAGERKURTH, C., STENZEL, R., & PRANTE, T. (2003). *STARS - a ubiquitous computing platform for computer augmented tabletop games.* Paper presented at Video Track of Ubiquitous Computing (UBICOMP'03), Seattle, Washington, USA.

[73] MALONE, T. W. (1982). Heuristics for designing enjoyable user interfaces: Lessons from computer games In *Proceedings of the 1982 conference on human factors in computing systems* (pp. 63-68). Gaithersburg, Maryland, United States ACM Press.

[74] MANDRYK, R. L. (2004). *Measuring and enhancing human experience with entertainment technology using psychophysiological techniques.* Burnaby, BC: Depth Paper: School of Computing Science, Simon Fraser University.

[75] MANDRYK, R. L., INKPEN, K. M., BILEZIKJIAN, M., KLEMMER, S. R., & LANDAY, J. A. (2001). *Supporting children's collaboration across handheld computers.* Short Paper presented at Conference on Human Factors in Computing Systems, Seattle, WA, USA., pp. 255-256.

[76] MANDRYK, R. L., MARANAN, D. S., & INKPEN, K. M. (2002). *False prophets: Exploring hybrid board/video games.* Short Paper presented at Conference on Human Factors in Computing Systems, pp. 640-641.

[77] MARCHIONINI, G. (1990). Evaluating hypermedia-based learning. In D. H. Jonassen & H. Mandl (Eds.), *Designing hypertext/hypermedia for learning* (pp. 355-373). Heidelberg, Germany: Springer-Verlag.

[78] MARRIN, T., & PICARD, R. W. (1998). *Analysis of affective musical expression with the conductor's jacket.* Paper presented at XII Colloquium on Musical Informatics, Gorizia, Italy.

[79] MARSHALL, C., & ROSSMAN, G. B. (1999). *Designing qualitative research* (3rd ed.). Thousand Oaks: Sage Publications.

[80] MARTINI, F. H., & TIMMONS, M. J. (1997). *Human anatomy* (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall.

[81] MAXWELL, J. A. (1996). *Qualitative research design: An interactive approach.* Thousand Oaks: Sage.

[82] MEDLOCK, M., WIXON, D., TERRANO, M., ROMERO, R., & FULTON, B. (2002). *Using the RITE method to improve products: A definition and a case study.* Paper presented at Usability Professionals Association, Orlando, FL.

[83] MESHKATI, N. (1988). Heart rate variability and mental workload assessment. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 101-115). North-Holland: Elsevier Science Publishers.

[84] MILLEN, D. R. (2000). *Rapid ethnography: Time deepening strategies for HCI field research.* Paper presented at Designing Interactive Systems, New York City, New York, pp. 280-286.

[85] *Mit media lab: Affective computing research.* Retrieved January 2004, from http://affect.media.mit.edu/

[86] MOLICH, R., & NIELSEN, J. (1990). Improving a human-computer dialogue. *Communications of the ACM, 33*(3), 338-348.

[87] MULDER, G. (1979). Sinusarrhythmia and mental workload. In N. Moray (Ed.), *Mental workload: Its theory and measurement* (pp. 299-325). New York: Plenum.

[88] MYRTEK, M., WEBER, D., BRÜGNER, G., & MÜLLER, W. (1996). Occupational stress and strain of female students: Results of physiological, behavioral, and psychological monitoring. *Biological Psychology, 42*, 379-391.

[89] NASOZ, F., OZYER, O., LISETTU, C. L., & FINKELSTEIN, N. (2002). *Multimodal affective driver interfaces for future cars.* Paper presented at Multimedia 2002, Juan-les-Pins, France, pp. 319-322.

[90] NIELSEN, J. (1992). Evaluating the thinking-aloud technique for use by computer scientists. In H. R. Hartson & D. Hix (Eds.), *Advances in human-computer interaction* (Vol. 3, pp. 69-82). Norwood: Ablex Publishing Corporation.

[91] NIELSEN, J. (1992). *Finding usability problems through heuristic evaluation.* Paper presented at Conference on Human Factors in Computing Systems, Monterey, CA, USA, pp. 373-380.

[92] NIELSEN, J. (1993). *Usability engineering*. Boston: Academic Press.

[93] NIELSEN, J. (1994). Heuristic evaluation. In J. Nielsen (Ed.), *Usability inspection methods* (pp. 25-62). New York: John Wiley and Sons, Inc.

[94] NORMAN, D. A. (2002, July 2002). Emotion and design: Attractive things work better. *Interactions, 9,* 36-42.

[95] OLSON, R. P. (1995). Definitions of biofeedback and applied psychophysiology. In M. S. Schwartz (Ed.), *Biofeedback: A practitioner's guide*. New York: The Guilford Press.

[96] PAGULAYAN, R. J., KEEKER, K., WIXON, D., ROMERO, R., & FULLER, T. (2002). User-centered design in games. In J. Jacko & A. Sears (Eds.), *Handbook for human-computer interaction in interactive systems* (pp. 883-906). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

[97] PAPILLO, J. F., & SHAPIRO, D. (1990). The cardiovascular system. In L. G. Tassinary (Ed.), *Principles of psychophysiology: Physical, social, and inferential elements* (pp. 456-512). Cambridge: Cambridge University Press.

[98] PARTALA, T., & SURAKKA, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with Computers, 16*, 295-309.

[99] PARTALA, T., SURAKKA, V., & VANHALA, T. (2005). Person-independent estimation of emotional experiences from facial expressions In *Proceedings of the 10th international conference on intelligent user interfaces* (pp. 246-248). San Diego, California, USA ACM Press.

[100] *Philips smart underwear.* Retrieved January 2004, from http://www.betterhumans.com/Errors/index.aspx?aspxerrorpath=/Search_Engine_Links/2003/searchEngineLink.article.2003-10-10-5.aspx

[101] PICARD, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

[102] PICARD, R. W., & HEALEY, J. (1997). *Affective wearables.* Paper presented at The First International Symposium on Wearable Computers.

[103] PIECHULLA, W., MAYSER, C., GEHRKE, H., & KÖNIG, W. (2003). Reducing drivers' mental workload by means of an adaptive man-machine interface. *Transportation Research, F*(6), 233-248.

[104] PINELLE, D., & GUTWIN, C. (2002). *Groupware walkthrough: Adding context to groupware usability evaluation.* Paper presented at Conference on Human Factors in Computing Systems, Minneapolis, MN, USA, pp. 455-462.

[105] PINELLE, D., & GUTWIN, C. (2003). Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM Transactions on Computer-Human Interaction (TOCHI), 10*(4), 281-311.

[106] POLSON, P., LEWIS, C., RIEMAN, J., & WHARTON, C. (1992). Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies, 36*, 741-773.

[107] RANI, P., SARKAR, N., SMITH, C. A., & KIRBY, L. D. (2004). Anxiety detecting robotic system- towards implicit human-robot collaboration. *Robotica, 22*, 85-95.

[108] RANI, P., SIMS, J., BRACKIN, R., & SARKAR, N. (2002). Online stress detection using psychophysiological signal for implicit human-robot cooperation. *Robotica, 20*(6), 673-686.

[109] RAU, R. (1996). Psychophysiological assessment of human reliability in a simulated complex system. *Biological Psychology, 42*, 287-300.

[110] RICHTER, P., WAGNER, T., HEGER, R., & WEISE, G. (1998). Psychophysiological analysis of mental load during driving on rural roads- a quasi-experimental field study. *Ergonomics, 41*(5), 593-609.

[111] RITTER, F. E., & LARKIN, J. H. (1994). Developing process models as summaries of HCI action sequences. *Human-Computer Interaction, 9*(3), 345-383.

[112] ROSENBERG, R. (1998). *Computing without mice and keyboards: Text and graphic input devices for mobile computing.* Unpublished Doctoral Dissertation, Department of Computer Science: University College, London.

[113] ROWE, D. W., SIBERT, J., & IRWIN, D. (1998). *Heart rate variability: Indicator of user state as an aid to human-computer interaction.* Paper presented at Conference on Human Factors in Computing Systems, pp. 480-487.

[114] RUSSELL, J. A., WEISS, A., & MENDELSOHN, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493-502.

[115] *S.M.A.R.T. Braingames.* Retrieved January 2004, from http://www.cyberlearningtechnology.com/

[116] SAMMER, G. (1998). Heart period variability and respiratory changes associated with physical and mental load: Non-linear analysis. *Ergonomics, 41*(5), 746-755.

[117] SANDERSON, P., & FISHER, C. (1994). Exploratory sequential data analysis: Foundations. *Human-Computer Interaction, 9*(3), 251-317.

[118] SCERBO, M. W., FREEMAN, F. G., MIKULKA, P. J., PARASURAMAN, R., DI NOCERO, F., & PRINZEL, L. J. (2001). *The efficacy of psychophysiological measures for implementing adaptive technology* (No. NASA/TP-2001-211018).

[119] SCOTT, S. D., MANDRYK, R. L., & INKPEN, K. M. (2002). *Understanding children's interactions in synchronous shared environments.* Paper presented at Computer Supported Collaborative Learning, Boulder, CO, USA, pp. 333-341.

[120] SCOTT, S. D., MANDRYK, R. L., & INKPEN, K. M. (2003). Understanding children's collaborative interactions in shared environments. *Journal of Computer Assisted Learning, 19*(2), 220-228.

[121] *Sensatex.* Retrieved January 2004, from http://www.sensatex.com/

[122] SHNEIDERMAN, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (Third ed.). Reading, MA: Addison, Wesley, Longman.

[123] STERN, R. M., RAY, W. J., & QUIGLEY, K. S. (2001). *Psychophysiological recording* (2nd ed.). New York: Oxford University Press.

[124] STROOP, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

[125] SURAKKA, V., ILLI, M., & ISOKOSKI, P. (2004). Gazing and frowning as a new human-computer interaction technique. *ACM Transactions on Applied Perceptions, 1*(1), 40-56.

[126] SWEENEY, M., MAGUIRE, M., & SHACKEL, B. (1993). Evaluating user-computer interaction: A framework. *International Journal of Man-Machine Studies, 38*, 689-711.

[127] SWEETSNER, P., & WYETH, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *ACM Computers in Entertainment, 3*(3), Article 3A.

[128] SYKES, J., & BROWN, S. (2003). *Affective gaming: Measuring emotion through the gamepad.* Short Paper presented at Conference on Human Factors in Computing Systems, Ft. Lauderdale, FA, USA, pp. 732-733.

[129] *Thought technologies cardiopro manual.* Retrieved September, 2005, from http://www.thoughttechnologies.com/pdf/SA7570-02.pdf

[130] *Thought technology.* Retrieved February 2004, from www.thoughttechnology.com

[131] TSOUKALAS, L. H., & UHRIG, R. E. (1997). *Fuzzy and neural approaches in engineering.* New York: John Wiley & Sons, Inc.

[132] TURKLE, S. (1995). *Life on the screen: Identity in the age of the internet.* New York: Touchstone.

[133] VAN RAVENSWAAIJ-ARTS, C. M. A., KOLLEE, L. A. A., HOPMAN, J. C. W., STOELINGA, G. B. A., & VAN GEIJN, H. P. (1993). Heart rate variability. *Annals of Internal Medicine, 118*(6), 436-447.

[134] VELTMAN, J. A., & GAILLARD, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology, 42*, 323-342.

[135] VELTMAN, J. A., & GAILLARD, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics, 41*(5), 656-669.

[136] VENABLES, P. H., & CHRISTIE, M. H. (1973). Mechanisms, instrumentation, recording techniques, and quantification of responses. In W. F. Prokasy & D. C. Raskin (Eds.), *Electrodermal activity in psychological research* (pp. 2-124). New York: Academic Press.

[137] VICENTE, K. J., THORNTON, D. C., & MORAY, N. (1987). Spectral analysis of sinus arrhythmia: A measure of mental effort. *Human Factors, 29*(2), 171-182.

[138] VORTAC, O. U., EDWARDS, M. B., & MANNIN, C. A. (1994). Sequences of actions for individual and teams of air traffic controllers. *Human-Computer Interaction, 9*(3), 319-343.

[139] WARD, R. D., & MARSDEN, P. H. (2003). Physiological responses to different web page designs. *International Journal of Human-Computer Studies, 59*(1/2), 199-212.

[140] WARD, R. D., MARSDEN, P. H., CAHILL, B., & JOHNSON, C. (2002). *Physiological responses to well-designed and poorly designed interfaces.* Paper presented at CHI 2002 Workshop on Physiological Computing, Minneapolis, MN, USA.

[141] WASTELL, D. G., & NEWMAN, M. (1996). Stress, control and computer system design: A psychophysiological field study. *Behaviour and Information Technology, 15*(3), 183-192.

[142] WHARTON, C., RIEMAN, J., LEWIS, C., & POLSON, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Nielsen (Ed.), *Usability inspection methods* (pp. 105-140). New York: John Wiley and Sons, Inc.

[143] WIENTJES, C. J. E. (1992). Respiration in psychophysiology: Methods and applications. *Biological Psychology, 34*, 179-203.

[144] WILSON, G. F. (1992). Applied use of cardiac and respiration measures: Practical considerations and precautions. *Biological Psychology, 34*, 163-178.

[145] WILSON, G. F. (2001). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology, 12*(1), 3-18.

[146] WILSON, G. F., LAMBERT, J. D., & RUSSELL, C. A. (1999). *Performance enhancement with real-time physiologically controlled adaptive aiding.* Paper presented at Human Factors and Ergonomics Society 44th Annual Meeting, pp. 361-364.

[147] WILSON, G. M. (2001). *Psychophysiological indicators of the impact of media quality on users.* Paper presented at CHI 2001 Doctoral Consortium, Seattle, WA, USA., pp. 95-96.

[148] WILSON, G. M., & SASSE, M. A. (2000). *Do users always know what's good for them? Utilizing physiological responses to assess media quality.* Paper presented at HCI 2000: People and Computers XIV - Usability or Else!, Sunderland, UK., pp. 327-339.

[149] WILSON, G. M., & SASSE, M. A. (2000). *Investigating the impact of audio degradations on users: Subjective vs. Objective assessment methods.* Paper presented at OZCHI 2000: Interfacing Reality in the New Millennium, Sydney, Australia., pp. 135-142.

[150] WINTON, W., PUTNAM, L., & KRAUSS, R. (1984). Facial and autonomic manifestations of the dimensional structure of emotion. *Journal of Experimental Social Psychology, 20*, 195-216.

[151] *Xeodesign*. Retrieved August 2005, from http://xeodesign.com/about.html

[152] ZHAI, S. (2003). What's in the eyes for attentive input. *Communications of the ACM, 46*(3), 34-39.

APPENDICES

Appendix 1 ABBREVIATIONS AND ACRONYMS

˚C: degrees Celsius

ACM: Associated of Computing Machinery

ADHD: Attention Deficit Hyperactivity Disorder

ANOVA: Analysis of Variance

ANS: Autonomic Nervous System

ATC: Air Traffic Control

AT:ST: Analysis Time to Sequence Time Ratio

AV space: Arousal-Valence Space

BP: Blood Pressure

BVP: Blood Volume Pulse

CD: Compact Disk

CNS: Central Nervous System

EA: Electronic Arts

ECG: Electrocardiography

EDR: Electrodermal Response

EEG: Electroencephalography

EKG: Electrocardiography

EMG: Electromyography

ESDA: Exploratory Sequential Data Analysis

FFT: Fast Fourier Transform

GSR: Galvanic Skin Response

HCI: Human Computer Interaction

HEP: Heuristic Evaluation for Playability

HP: Heart Period

HSD: Honestly Significant Difference

HR: Heart Rate

HRV: Heart Rate Variability

Hz: Hertz (unit of frequency)

IAPS: International Affective Picture System

IBI: Inter-beat Interval

ID: Identification

LAN: Local Area Network

MANOVA: Multivariate Analysis of Variance

MIT: Massachusetts Institute of Technology

NASA: National Aeronautics and Space Administration

NHL: National Hockey League

***ns*:** non-significant

PDA: Personal Digital Assistant

PNS: Parasympathetic Nervous System

PS2: PlayStation 2

RespAmp: Respiration Amplitude

RespRate: Respiration Rate

RITE: Rapid Iterative Testing and Evaluation

RSA: Respiratory Sinus Arrhythmia

SC: Skin Conductance

SD: Standard Deviation (also St. Dev.)

SE: Standard Error

SNS: Somatic Nervous System/Sympathetic Nervous System

SRR: Skin Resistance Response

V_{MIN}: Minute Volume

V_T: Tidal Volume

Appendix 2 FALSE PROPHETS

Board games are highly interactive, provide a non-oriented interface, are mobile, and allow for a dynamic number of players and house rules. They also are limited to a fairly static environment, don't allow players to save the game state, and have simple scoring rules. On the other hand, computer games provide complex simulations, impartial judging, evolving environments, suspension of disbelief, and the ability to save game state. But computer games often support interaction with the system, rather than with other players. Even in a co-located environment, players sit side-by-side and interact with each other through the interface. The goal of developing a hybrid game system was to leverage the advantages of both of these mediums, encouraging interaction between the players. In False Prophets, players use tangible pieces to move around a digital game board, projected onto a table. The playing pieces are equipped with a button to perform simple game operations, while more complex interactions and private information is managed through a handheld computer. This unique game environment has the computational advantages of a computer game environment, while still supporting interpersonal interactions. In addition, it allowed for the development of novel game elements that couldn't exist with either of the traditional game technologies.

The Game

We created a hybrid platform to investigate this new class of games. Our game

environment consisted of a tabletop display system with a custom sensor interface. Initially, we configured the game for six players although the goal was to have dynamically changing groups. The game board was a projected map, tessellated into a grid of 20 by 30 hexagons. Each hexagon represented a space that the characters were allowed to occupy and was one of four terrain types: water, plains, forest, and mountains. Initially, the map was not projected, with the exception of hexagons where players were located. As the players moved around the board, the map was dynamically revealed. The players were separated into two teams and were initially unaware of their team members. The goal of the game was to discover which team each player belonged to. This was accomplished by gathering virtual clues, making virtual observations of the other players, and using this information to solve a logic puzzle. To support interpersonal interactions our rules encouraged players to concurrently and physically move around the board while communicating with each other in a face-to-face verbal or non-verbal exchange. We accomplished this through a number of game features.

Players gathered clues about others by physically moving their character around the game board, collecting clues that remained hidden in clue holders like rocks and logs. Players made observations by physically passing near other players on the game board. The level of detail of virtual observations (height vs. freckles) depended on the physical proximity of the playing pieces.

Private communication such as the exchange of clues and observations was not supported or mediated by the game. Any bargaining or player alliances had to occur between players in the physical world.

To avoid a static turn-taking strategy, which would not support interactivity, we implemented an energy-based system to move around the board. Each type of terrain had an associated energy factor that depleted the player's energy as they moved around the board. The characters' energy was replenished cyclically throughout the game and they had to time their explorations accordingly.

The Sensor Interface

To support players moving their characters around the projected display, we implemented a custom sensor interface. The playing surface contained an array of infrared phototransistors, each corresponding to a hexagon in the game. Each character playing piece contained an infrared light emitting diode. The pieces emitted a pulse that was sent though the phototransistors to the serial port and interpreted by the game software. Pieces also had buttons that were pressed to correspond to actions in the game. Pressing a button changed the pulse transmitted to the game. The pieces were a natural interface for players accustomed to dealing with physical figurines, yet provided a great deal of interactive functionality. These pieces, combined with the sensor array, provided us with seamless input to the game system. By making interaction with the computer components of the game seamless, we allowed players to focus on each other, and not on the interface.

The Handheld Interface

The display system consisted of both the tabletop projection for public information as well as handheld computers for private information. The handheld computers also acted as input to the game by allowing players to perform actions and make choices that could not be communicated naturally via the game pieces. We deliberately limited the interaction through the handhelds to maintain focus on the other players, not on the private displays. The handhelds communicated to the game control through an 802.11 wireless network. All public input occurred through the pieces, which connected to the game control via the serial port. The game control handled all game input, logic, and updated the display based on events in the game.

Appendix 3 GSR ELECTRODE PLACEMENT TESTS

Before beginning Experiment One, we ran a number of electrode placement tests to see whether electrode placement affected the GSR signal. One subject had the electrodes placed in two different locations and watched a video clip intended to create arousing and relaxing experiences. Electrodes were tested on the fingers, palm of the hand, and sole of the foot using big and small electrodes. The size of the metal contact in the large electrode was the same as the size of the metal contact in the small electrode; however, the size of the surrounding sticker that attaches the electrode to the skin was larger in the big electrode. The following locations were tested:

Finger clips: Two electrodes were attached to the index and ring finger using elastic and Velcro finger clips.

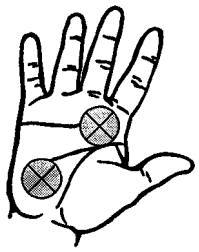
Feet: big electrodes: Two large electrodes were placed on the sole of the foot in the following configuration:

Feet: Pinky and Big Toe, small electrodes: Two small electrodes were placed on the sole of the foot in the following configuration:

Feet: Big Toe down and across, big electrodes: Two small electrodes were placed on the sole of the foot in the following configuration:

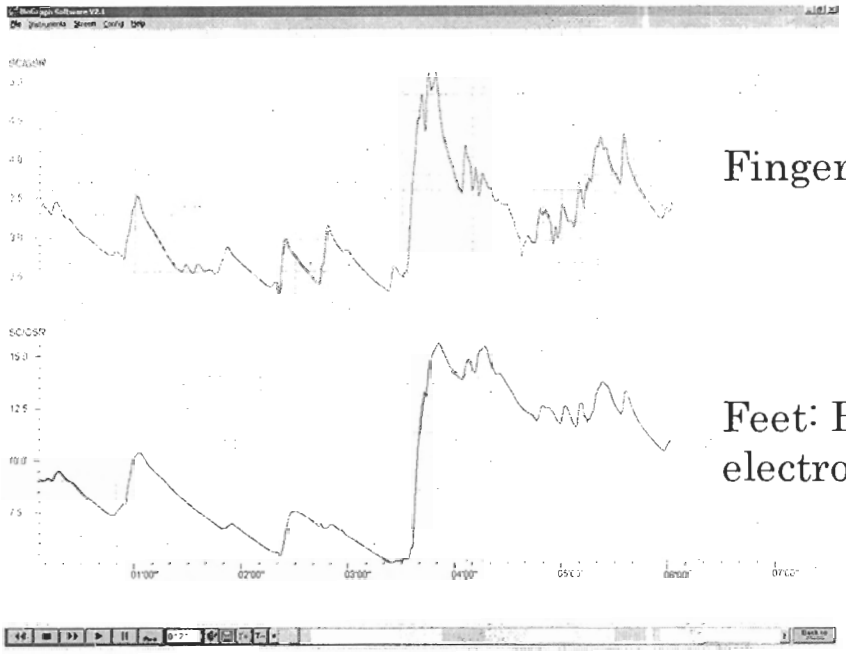
Palm: Pinky and thumb, small electrodes: Two small electrodes were placed on the palm of the hand in the following configuration:

Palm: index, down and across, big electrodes: Two big electrodes were placed on the palm of the hand in the following configuration:



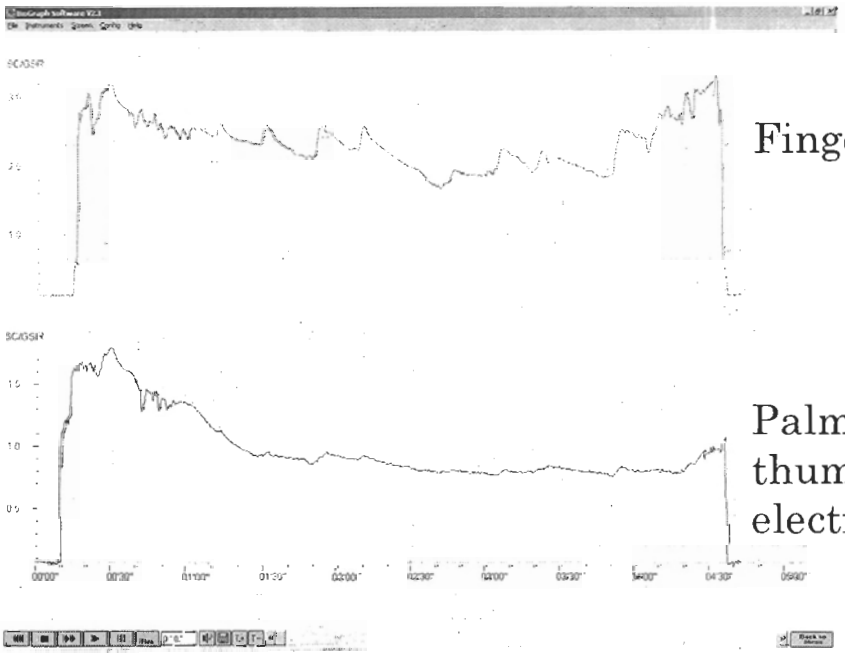
The results showed that using big electrodes on the feet produced the best (least noisy) signal; however, the finger clips were just as responsive (although on a different absolute μm scale), as the feet electrodes. Asking participants to remove their shoes and socks may have made them uncomfortable, thus the finger clips were judged less invasive than electrodes on the feet. Although the signal was slightly noisier, we decided that we could easily filter the GSR signal generated from the finger clips.

The following graphs show the output of the tests.



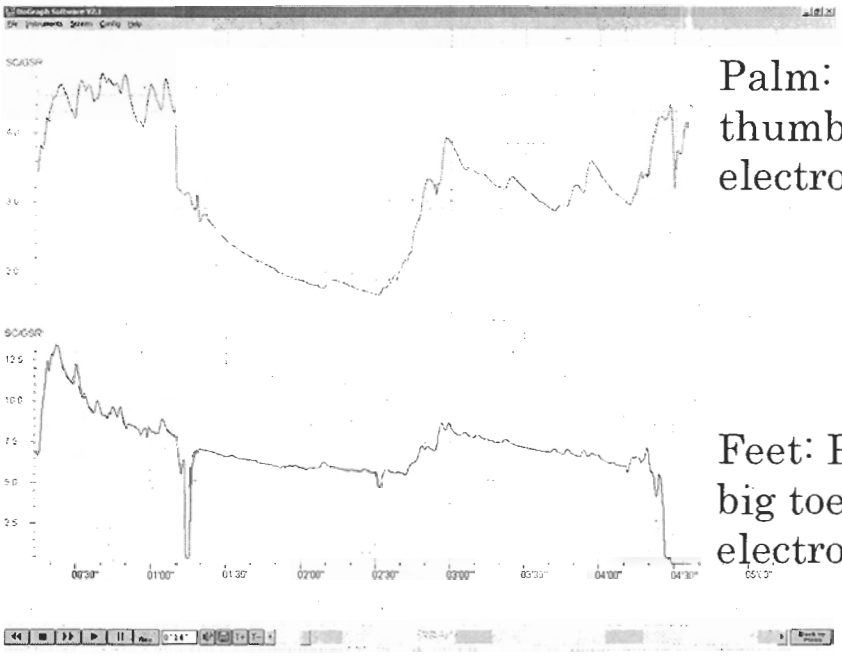
Finger clips

Feet: Big electrodes

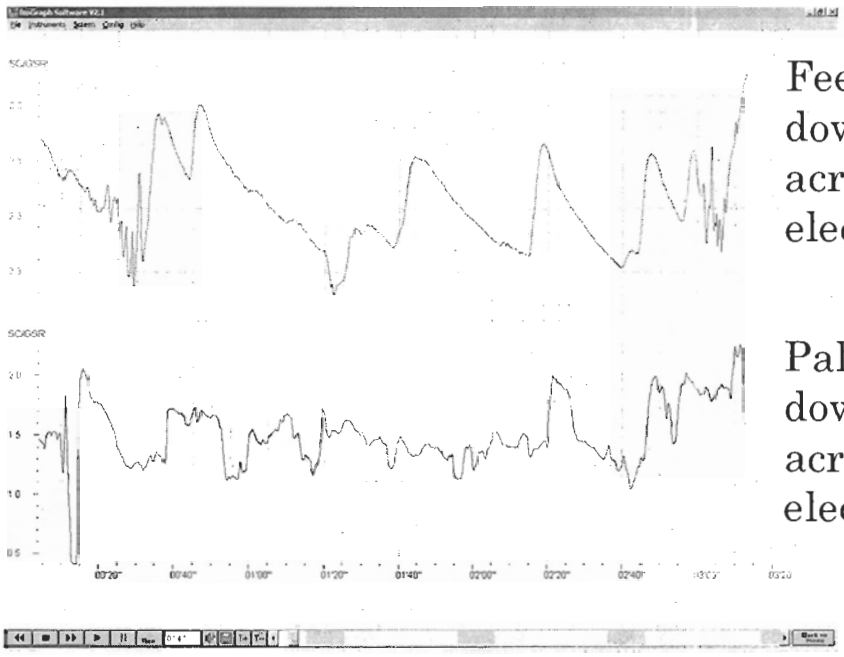


Finger clips

Palm: Pinky and thumb, small electrodes



We also observed the GSR signal generated by playing NHL2003 to see if hand movements impacted the signal. Electrodes on the palm did not work well, since movement when operating a game controller caused slight variations in the connectivity between the electrode and the skin, impacting the conductivity of the connection. Movements made when operating a game controller did not impact electrodes on the feet and the finger clips. An example of a test on the foot and palm follows:



Feet: big toe, down and across, big electrodes

Palm: index, down and across, big electrodes

Appendix 4 CONSENT FORM (ALL EXPERIMENTS)

SIMON FRASER UNIVERSITY

Form 2- Informed Consent By Participants In a Research Project or Experiment

The University and those conducting this project subscribe to the ethical conduct of research and to the protection at all times of the interests, comfort, and safety of subjects. This research is being conducted under permission of the Simon Fraser Research Ethics Board. The chief concern of the Board is for the health, safety and psychological well-being of research participants

Should you wish to obtain information about your rights as a participant in research, or about the responsibilities of researchers, or if you have any questions, concerns or complaints about the manner in which you were treated in this study, please contact the Director, Office of Research Ethics by email at hweinber@sfu.ca or phone at 604-268-6593.

Your signature on this form will signify that you have received a document which describes the procedures, possible risks, and benefits of this research project, that you have received an adequate opportunity to consider the information in the documents describing the project or experiment, and that you voluntarily agree to participate in the project or experiment.

Any information that is obtained during this study will be kept confidential to the full extent permitted by the law. Knowledge of your identity is not required. You will not be required to write your name on any other identifying information on research materials. Materials will be maintained in a secure location.

Title: **Measuring enjoyment of computer games using psychophysiological techniques**
Investigator Name: **Regan Mandryk**
Investigator Department: **Computing Science**

Having been asked to participate in a research project or experiment, I certify that I have read the procedures specified in the information documents, describing the project or experiment. I understand the procedures to be used in this experiment and the personal risks to me in taking part in the project or experiment, as stated below:

Risks and Benefits:
There are no risks involved. You will receive a monetary stipend or a product from EA Sports for participating in this study. There are no direct benefits to you however, the results of this research may contribute to the knowledge base of Human-Computer Interaction research and also may lead to the development of better game interfaces and game design techniques.

I understand that I may withdraw my participation at any time. I also understand that I may register any complaint with the Director of the Office of Research Ethics or the researcher named above or with the Chair, Director or Dean of the Department, School or Faculty as shown below.

Department, School or Faculty: **Computing Science**
Chair, Director or Dean: **Dr. Ze-Nian Li**

8888 University Way, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada

I may obtain copies of the results of this study, upon its completion by contacting:
Regan Mandryk or Dr. Tom Calvert at Simon Fraser University, 8888 University Way, Burnaby, BC, V5A 1S6, Canada

I have been informed that the research will be confidential.

I understand that my participation in my study of this kind may require me to obtain his or her permission prior to my participation in my study of this kind.

What The Subject is Required to Do:

9-30 year old individuals drawn from the University population through word of mouth and e-mail recruitement. Your task will be to play computer games on a the library of titles published by EA Sports. Game scenarios (e.g. snowboard race down a mountain) will be replayed to you once you have finished them. As these scenarios are replayed, you will be asked to verbally relate the experience you had while completing that game scenario. During the session, video cameras will record your physical actions with the game controller, your facial expressions, and any verbal communication, and a researcher will observe and take notes regarding your interactions with the console gaming system. Your physiological responses will be monitored using biometric equipment provided by Thought Technologies Inc. These responses include EKG (electrocardiography), EMG (electromyography) of the jaw, respiration, and GSR (galvanic skin response). EKG will be sensed using three surface electrodes on either the shoulders/abdomen or on the forearm. EMG will be sensed using three surface electrodes on the jaw. Sensing respiration involves using a Velcro and rubber strap wrapped around the chest, while GSR involves using two surface electrodes on either the hands or feet. You will be asked to complete a questionnaire before the experiment. This background questionnaire will help us determine your familiarity with computers, gaming systems and computer games. After the experiment, an informal interview will gather your opinions about playing the different games in the different conditions. You are allowed to withdraw your participation in the experiment at any time.

The subject and witness shall fill in this box. (Please Print Legibly)

Subject Last Name: Subject First Name:

Subject Contact Information:

Subject Signature: Witness:

Date (use format MM/DD/YYYY):

Appendix 5 BACKGROUND QUESTIONNAIRE

All of the following questionnaires were administered online.

Background Questionnaire

A. Personal Information

1. Name (optional):

2. Age:

3. Sex:

- ○ Male
- ○ Female

4. Handedness:

- ○ Right
- ○ Left

B. Computer/Console Usage

5. How often do you:

	Never	Rarely	Occasionally	Often	Every day
Use computers?	○	○	○	○	○
Play computer games?	○	○	○	○	○
Play video (console) games?	○	○	○	○	○
Play computer/video games over the internet or network?	○	○	○	○	○
Play computer/video games with another co-located player?	○	○	○	○	○

6. How much time do you spend:

	None	Less than 3 hours a week	3 - 7 hours a week	1 - 2 hours a day	More than 2 hours a day
Using computers?	○	○	○	○	○
Playing video games?	○	○	○	○	○
Playing video (console) games?	○	○	○	○	○
Playing computer/video games over the internet or a network?	○	○	○	○	○
Playing computer/video games with another co-located player?	○	○	○	○	○

7. Which gaming systems do you use? (check all that apply):

- ☐ Computer
- ☐ Game Boy
- ☐ Game Gear
- ☐ Handheld Computer
- ☐ Cell Phone
- ☐ Nintendo
- ☐ Super Nintendo
- ☐ Nintendo 64
- ☐ Nintendo Game Cube
- ☐ Sega CD
- ☐ Sega Saturn
- ☐ Sega Dreamcast
- ☐ Sony Playstation
- ☐ Sony PS2
- ☐ X-Box
- ☐ Arcade
- ☐ Other

8. Which gaming systems do you own? (check all that apply):

- ☐ Computer
- ☐ Game Boy
- ☐ Game Gear
- ☐ Handheld Computer
- ☐ Cell Phone
- ☐ Nintendo
- ☐ Super Nintendo
- ☐ Nintendo 64
- ☐ Nintendo Game Cube
- ☐ Sega CD
- ☐ Sega Saturn
- ☐ Sega Dreamcast
- ☐ Sony Playstation
- ☐ Sony PS2
- ☐ X-Box
- ☐ Arcade
- ☐ Other

C. Game Preference

9. How much do you like each game genre (not including arcade games):

	Dislike a lot	Dislike somewhat	Neutral	Like Somewhat	Like a lot
Action	○	○	○	○	○
Adventure	○	○	○	○	○
Puzzle	○	○	○	○	○
Racing	○	○	○	○	○
Roleplaying	○	○	○	○	○
Shooting	○	○	○	○	○
Simulation	○	○	○	○	○
Sports	○	○	○	○	○
Strategy	○	○	○	○	○

10. List your favorite games:

11. List any games that you have found to be boring:

12. Have you ever chosen not to learn a game because it was too frustrating to play?

○ Yes
○ No

12a. Which games?

13. When you have been really into a game, how long have you continuously played a game?

14. Rate how experienced you are with the following game titles (or their previous versions):

	Very experienced	Somewhat experienced	Neutral	Somewhat inexperienced	Very inexperienced
NHL 2004 Hockey	○	○	○	○	○
SSX Snowboarding	○	○	○	○	○

Submit Survey

Powered by Perseus SurveySolutions.

Appendix 6 GOLDILOCKS CONDITION QUESTIONNAIRE

Date:_____
Time:_____

Participant ID:_____

Level of Play:	Beginner	Easy	Medium	Hard
Order: BEMD	EMDB	MDBE	DBEM	
DMEB	MEBD	EBDM	BDME	

Beginner	low				high	
Boredom:	0	1	2	3	4	_____
Frustration:	0	1	2	3	4	_____
Challenge:	0	1	2	3	4	_____
Fun:	0	1	2	3	4	_____

Easy	low				high	
Boredom:	0	1	2	3	4	_____
Frustration:	0	1	2	3	4	_____
Challenge:	0	1	2	3	4	_____
Fun:	0	1	2	3	4	_____

Medium	low				high	
Boredom:	0	1	2	3	4	_____
Frustration:	0	1	2	3	4	_____
Challenge:	0	1	2	3	4	_____
Fun:	0	1	2	3	4	_____

Hard	low				high	
Boredom:	0	1	2	3	4	_____
Frustration:	0	1	2	3	4	_____
Challenge:	0	1	2	3	4	_____
Fun:	0	1	2	3	4	_____

Appendix 7 GOLDILOCKS: POST EXPERIMENT QUESTIONNAIRE

Rankings:

Challenge: ___ Beginner ___ Easy ___ Medium ___ Hard

Explanation:_____

_____

Excitement: ___ Beginner ___ Easy ___ Medium ___ Hard

Explanation:_____

_____

Fun: ___ Beginner ___ Easy ___ Medium ___ Hard

Explanation:_____

_____

Comments and notes:

Appendix 8 TURING CONDITION QUESTIONNAIRE

Condition Questionnaire

1.

Name or ID

2. Level of Play

- Easy
- Beginner
- Medium
- Hard

3. Condition

- Against the Computer
- Against my Partner

4. The final score was:

Me

My partner

5. Please rate whether you agree or disagree with the following statements. This condition was _____ (insert words below).

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Boring					
Challenging					
Easy					
Engaging					
Exciting					
Frustrating					
Fun					

6. Any comments?

Submit Survey

Appendix 9 TURING POST-EXPERIMENT QUESTIONNAIRE

PostExperiment Questionnaire

1.

ID

2. Choose One

	Partner	Computer
Enjoy	○	○
Fun	○	○
Challenging	○	○
Exciting	○	○

3. Explain

4. If you could choose to play against a friend or against the computer, which would you choose?

○ Friend
○ Computer

5. Why?

Submit Survey

Appendix 10 EXP. 3: CONDITION QUESTIONNAIRE

Condition Questionnaire

1.

Name or ID

2. Level of Play

- Easy
- Beginner
- Medium
- Hard

3. Condition

- Against the Computer
- Against a Friend
- Against a Stranger

4. The final score was:

Me

My partner

5. Please rate whether you agree or disagree with the following statements. This condition was _____ (insert words below).

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Boring					
Challenging					
Easy					
Engaging					
Exciting					
Frustrating					
Fun					

6. Any comments?

Submit Survey

Appendix 11 EXP. 3: POST EXP. QUESTIONNAIRE

Post Experiment Questionnaire

1.

ID

2. Choose which condition you found to be the most:

	Friend	Stranger	Computer
Enjoyable	○	○	○
Fun	○	○	○
Challenging	○	○	○
Exciting	○	○	○

3. Explain

4. If you could choose to play against a friend, stranger, or against the computer, which would you choose?

- ○ Friend
- ○ Stranger
- ○ Computer

5. Why?

Submit Survey

Appendix 12 RULES FOR TRANSFORMING PHYSIOLOGICAL SIGNALS INTO AROUSAL AND VALENCE

The following 22 rules were used in concert with the membership functions described in section 8.3.1 to convert GSR, HR, $EMG_{smiling}$, and $EMG_{frowning}$ into arousal and valence:

1. If (GSR is high) then (arousal is high)
2. If (GSR is mid-high) then (arousal is mid-high)
3. If (GSR is mid-low) then (arousal is mid-low)
4. If (GSR is low) then (arousal is low)
5. If (HR is low) then (arousal is low)
6. If (HR is high) then (arousal is high)
7. If (GSR is low) and (HR is high) then (arousal is mid-low)
8. If (GSR is high) and (HR is low) then (arousal is mid-high)
9. If (EMG_{frown} is high) then (valence is very low)
10. If (EMG_{frown} is mid) then (valence is low)
11. If (EMG_{smile} is mid) then (valence is high)
12. If (EMG_{smile} is high) then (valence is very high)
13. If (EMG_{smile} is low) and (EMG_{frown} is low) then (valence is neutral)
14. If (EMG_{smile} is high) and (EMG_{frown} is low) then (valence is very high)
15. If (EMG_{smile} is high) and (EMG_{frown} is mid) then (valence is high)
16. If (EMG_{smile} is low) and (EMG_{frown} is high) then (valence is very low)
17. If (EMG_{smile} is mid) and (EMG_{frown} is high) then (valence is low)
18. If (EMG_{smile} is low) and (EMG_{frown} is low) and (HR is low) then (valence is low)
19. If (EMG_{smile} is low) and (EMG_{frown} is low) and (HR is high) then (valence is high)
20. If (GSR is high) and (HR is mid) then (arousal is high)
21. If (GSR is mid-high) and (HR is mid) then (arousal is mid-high)
22. If (GSR is mid-low) and (HR is mid) then (arousal is mid-low)

Appendix 13 Rules for Transforming Arousal and Valence into Five Emotional States

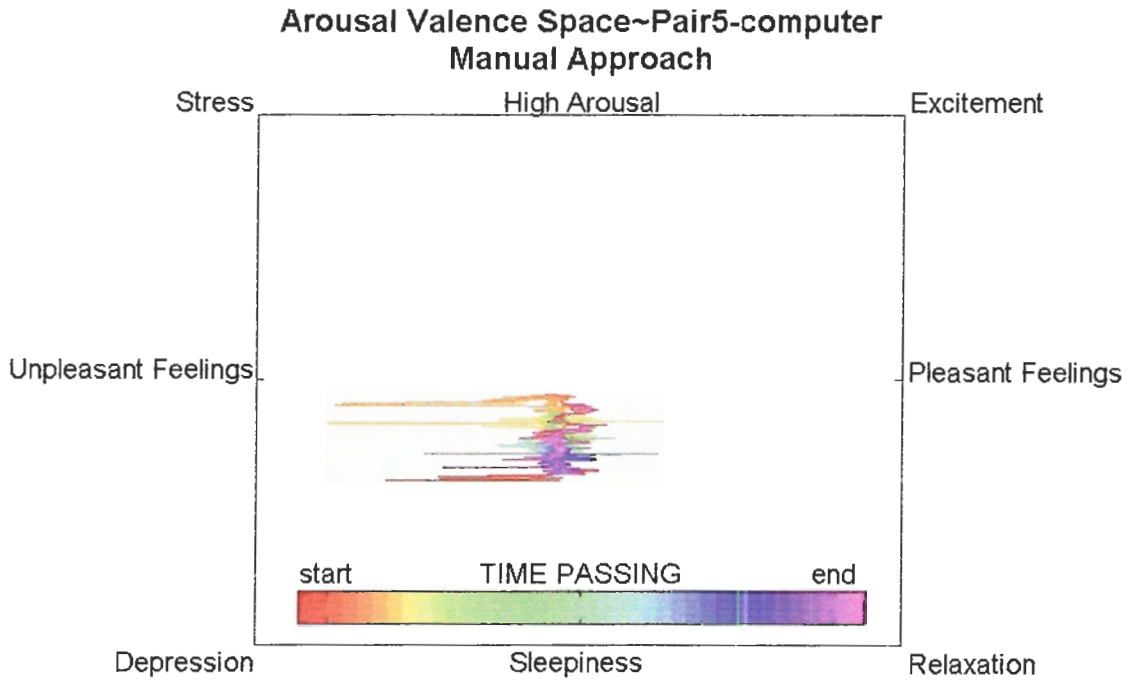
The following 67 rules were used in concert with the membership functions described in section 8.4.1 to convert arousal and valence into boredom, challenge, excitement, frustration, and fun:

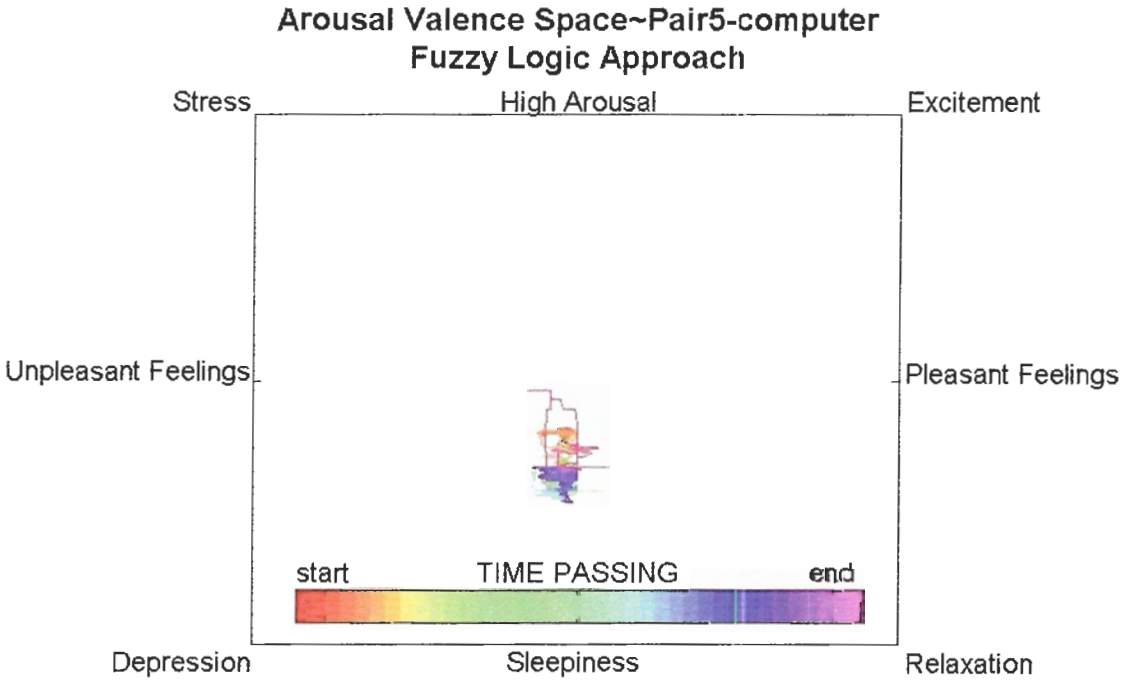
1. If (arousal is not veryLow) and (valence is midHigh) then (fun is low)
2. If (arousal is not low) and (valence is midHigh) then (fun is low)
3. If (arousal is not veryLow) and (valence is high) then (fun is medium)
4. If (valence is veryHigh) then (fun is high)
5. If (arousal is midHigh) and (valence is midLow) then (challenge is low)
6. If (arousal is midHigh) and (valence is midHigh) then (challenge is low)
7. If (arousal is high) and (valence is midLow) then (challenge is medium)
8. If (arousal is high) and (valence is midHigh) then (challenge is medium)
9. If (arousal is veryHigh) and (valence is midLow) then (challenge is high)
10. If (arousal is veryHigh) and (valence is midHigh) then (challenge is high)
11. If (arousal is midLow) and (valence is midLow) then (boredom is low)
12. If (arousal is midLow) and (valence is low) then (boredom is medium)
13. If (arousal is low) and (valence is low) then (boredom is medium)
14. If (arousal is low) and (valence is midLow) then (boredom is medium)
15. If (arousal is midLow) and (valence is veryLow) then (boredom is high)
16. If (arousal is low) and (valence is veryLow) then (boredom is high)
17. If (arousal is veryLow) and (valence is veryLow) then (boredom is high)
18. If (arousal is veryLow) and (valence is low) then (boredom is high)
19. If (arousal is veryLow) and (valence is midLow) then (boredom is high)
20. If (arousal is midHigh) and (valence is midLow) then (frustration is low)
21. If (arousal is midHigh) and (valence is low) then (frustration is medium)
22. If (arousal is high) and (valence is low) then (frustration is medium)
23. If (arousal is high) and (valence is midLow) then (frustration is medium)

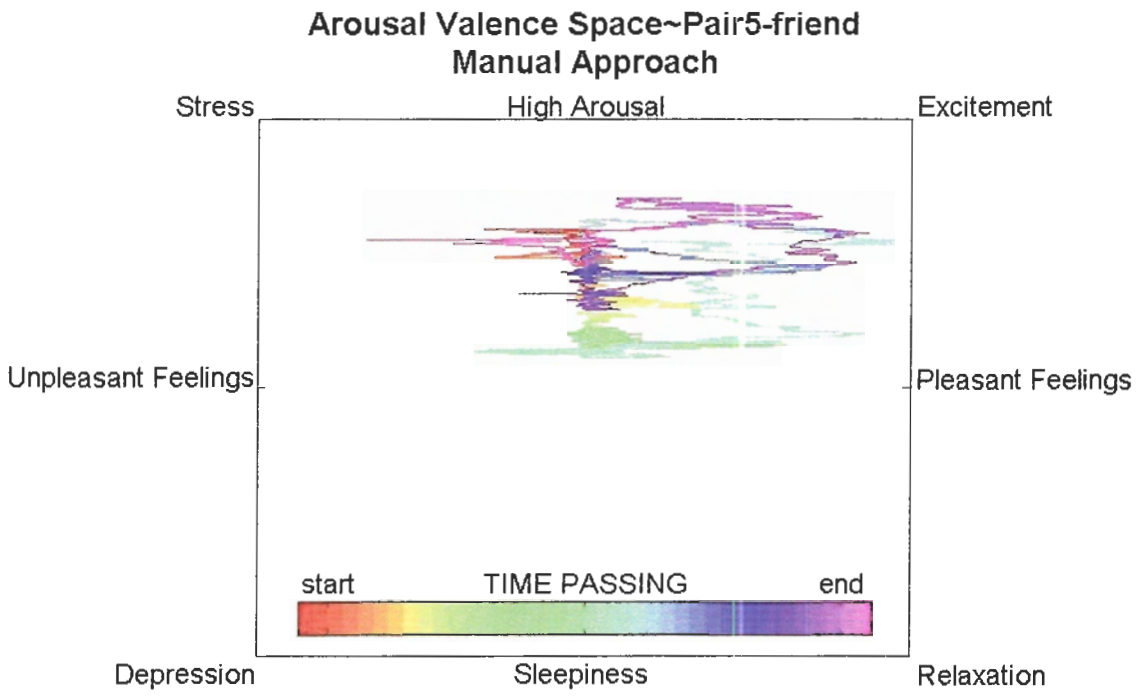
24. If (arousal is midHigh) and (valence is veryLow) then (frustration is high)
25. If (arousal is high) and (valence is veryLow) then (frustration is high)
26. If (arousal is veryHigh) and (valence is veryLow) then (frustration is high)
27. If (arousal is veryHigh) and (valence is low) then (frustration is high)
28. If (arousal is veryHigh) and (valence is midLow) then (frustration is high)
29. If (valence is veryLow) then (fun is veryLow)(challenge is veryLow)
30. If (valence is low) then (fun is veryLow)(challenge is veryLow)
31. If (valence is high) then (challenge is veryLow)(boredom is veryLow)(frustration is veryLow)
32. If (valence is veryHigh) then (challenge is veryLow)(boredom is veryLow)(frustration is veryLow)
33. If (valence is midHigh) then (boredom is veryLow)(frustration is veryLow)
34. If (arousal is veryLow) then (challenge is veryLow)(frustration is veryLow)
35. If (arousal is low) then (challenge is veryLow)(frustration is veryLow)
36. If (arousal is midLow) then (challenge is veryLow)(frustration is veryLow)
37. If (arousal is midHigh) then (boredom is veryLow)
38. If (arousal is high) then (boredom is veryLow)
39. If (arousal is veryHigh) then (boredom is veryLow)
40. If (arousal is veryLow) and (valence is midHigh) then (fun is veryLow)
41. If (arousal is low) and (valence is midHigh) then (fun is veryLow)
42. If (arousal is veryLow) and (valence is high) then (fun is low)
43. If (valence is midLow) then (fun is veryLow)
44. If (arousal is veryLow) and (valence is high) then (boredom is low)
45. If (arousal is low) and (valence is midHigh) then (boredom is low)
46. If (arousal is veryLow) and (valence is midHigh) then (boredom is medium)
47. If (arousal is veryHigh) and (valence is veryLow) then (challenge is medium)
48. If (arousal is veryHigh) and (valence is veryHigh) then (challenge is medium)
49. If (arousal is high) and (valence is low) then (challenge is low)
50. If (arousal is high) and (valence is high) then (challenge is low)
51. If (arousal is veryHigh) and (valence is low) then (challenge is high)
52. If (arousal is veryHigh) and (valence is high) then (challenge is high)
53. If (arousal is midHigh) and (valence is midHigh) then (excitement is low)
54. If (arousal is high) and (valence is midHigh) then (excitement is medium)
55. If (arousal is high) and (valence is high) then (excitement is medium)
56. If (arousal is midHigh) and (valence is high) then (excitement is medium)
57. If (arousal is veryHigh) and (valence is midHigh) then (excitement is high)

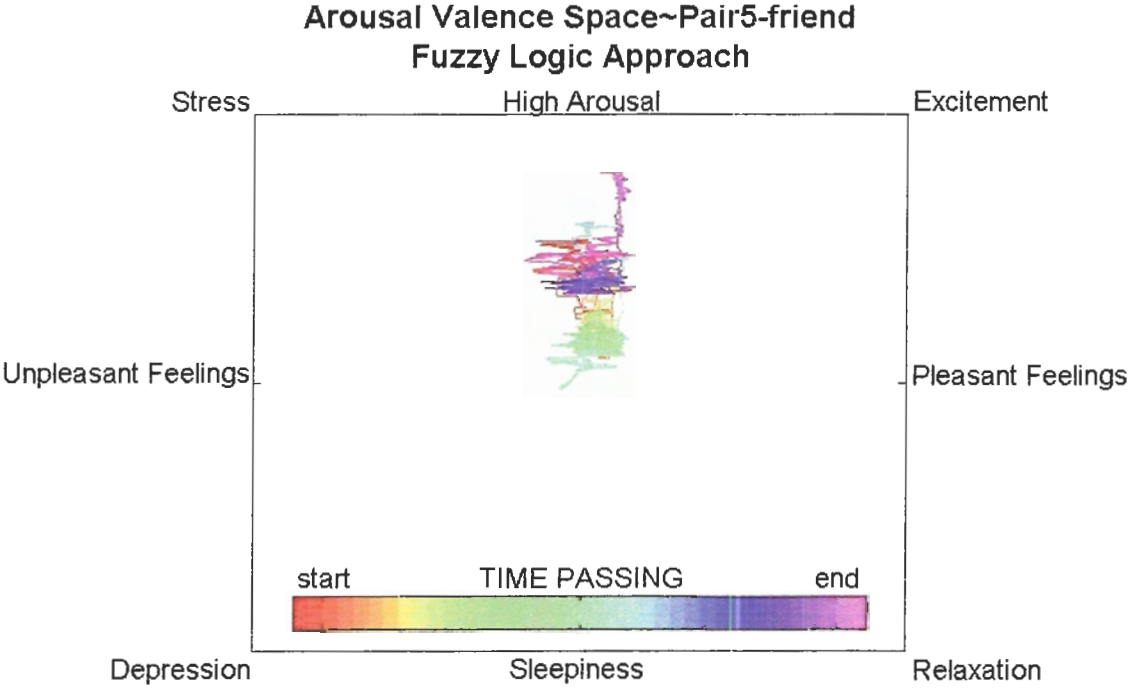
58. If (arousal is veryHigh) and (valence is high) then (excitement is high)
59. If (arousal is veryHigh) and (valence is veryHigh) then (excitement is high)
60. If (arousal is high) and (valence is veryHigh) then (excitement is high)
61. If (arousal is midHigh) and (valence is veryHigh) then (excitement is high)
62. If (arousal is midLow) then (excitement is veryLow)
63. If (arousal is low) then (excitement is veryLow)
64. If (arousal is veryLow) then (excitement is veryLow)
65. If (valence is veryLow) then (excitement is veryLow)
66. If (valence is low) then (excitement is veryLow)
67. If (valence is midLow) then (excitement is veryLow)

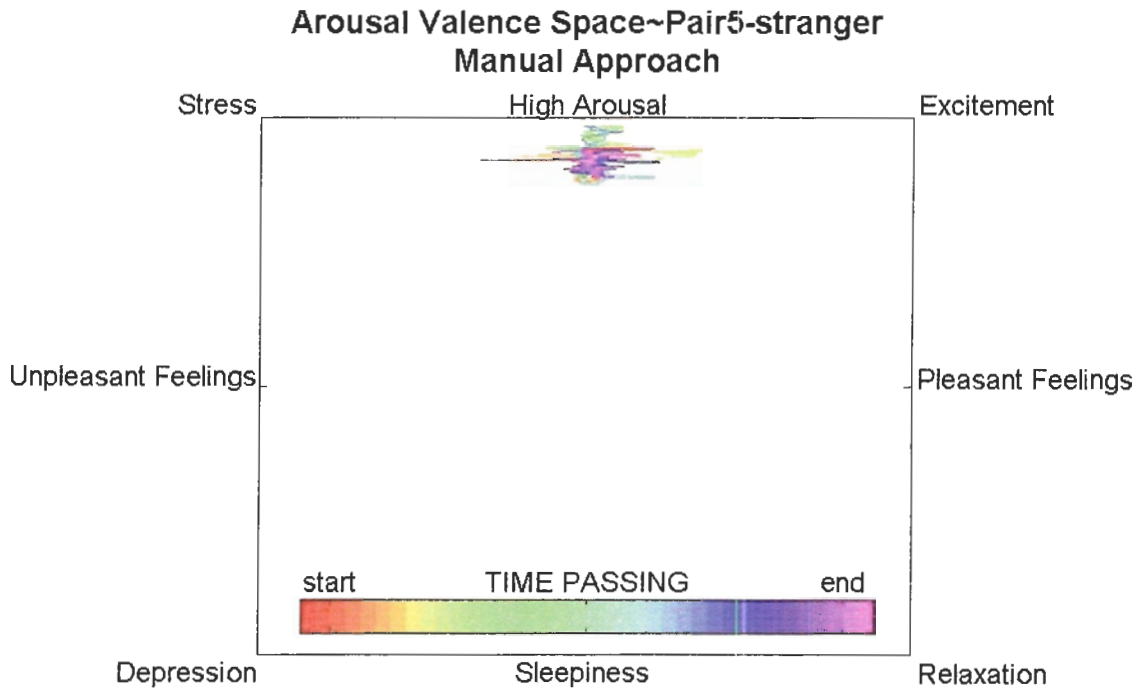
Appendix 14 EXP 3: AV SPACE GRAPHS

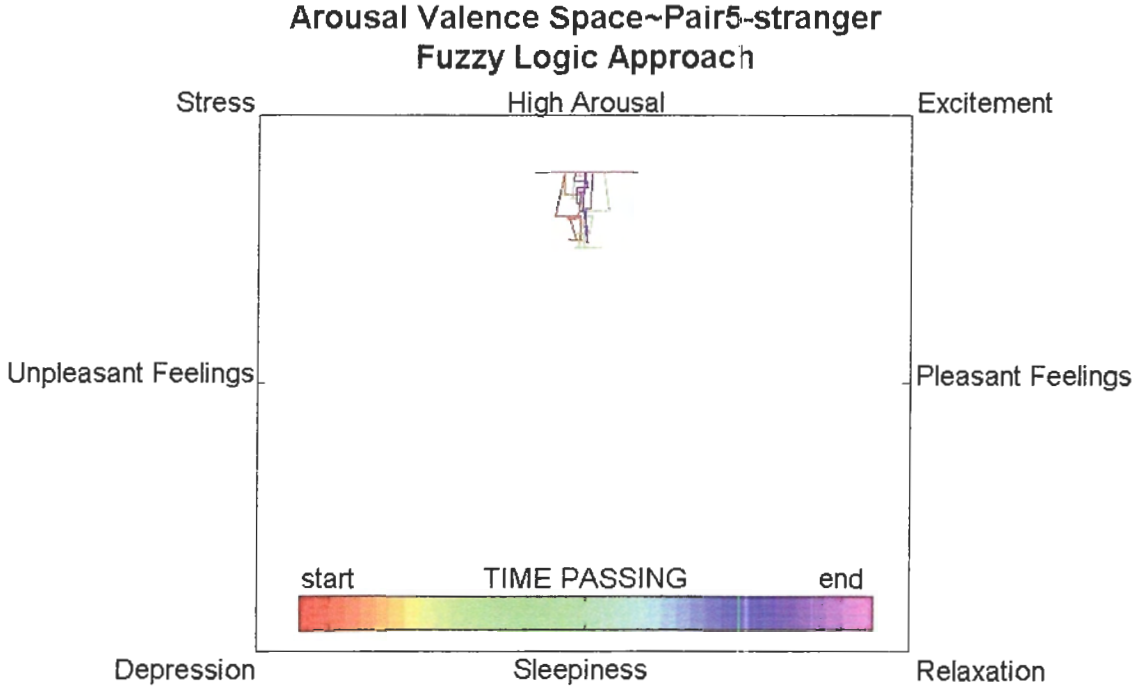


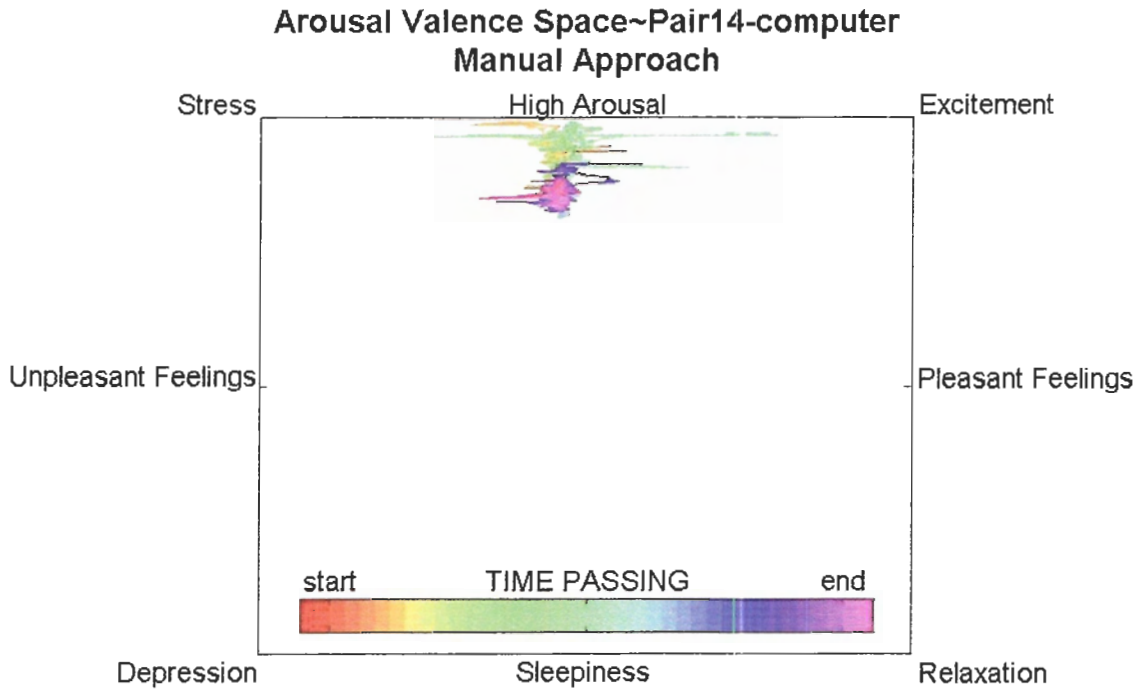


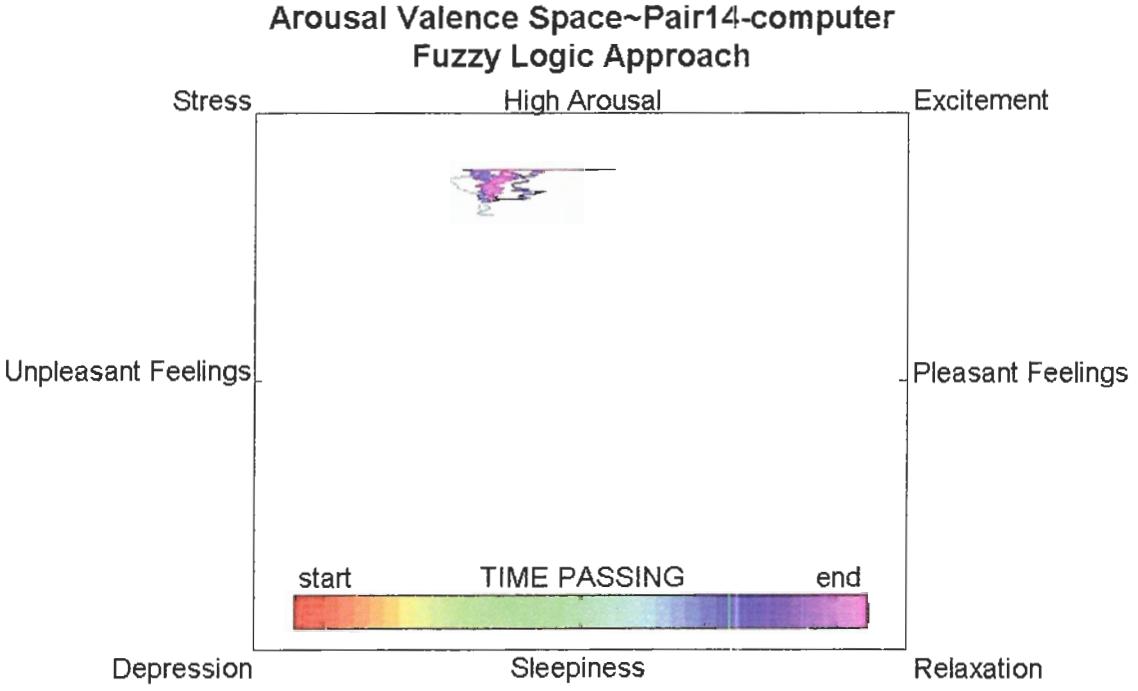

Arousal Valence Space~Pair5-friend
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

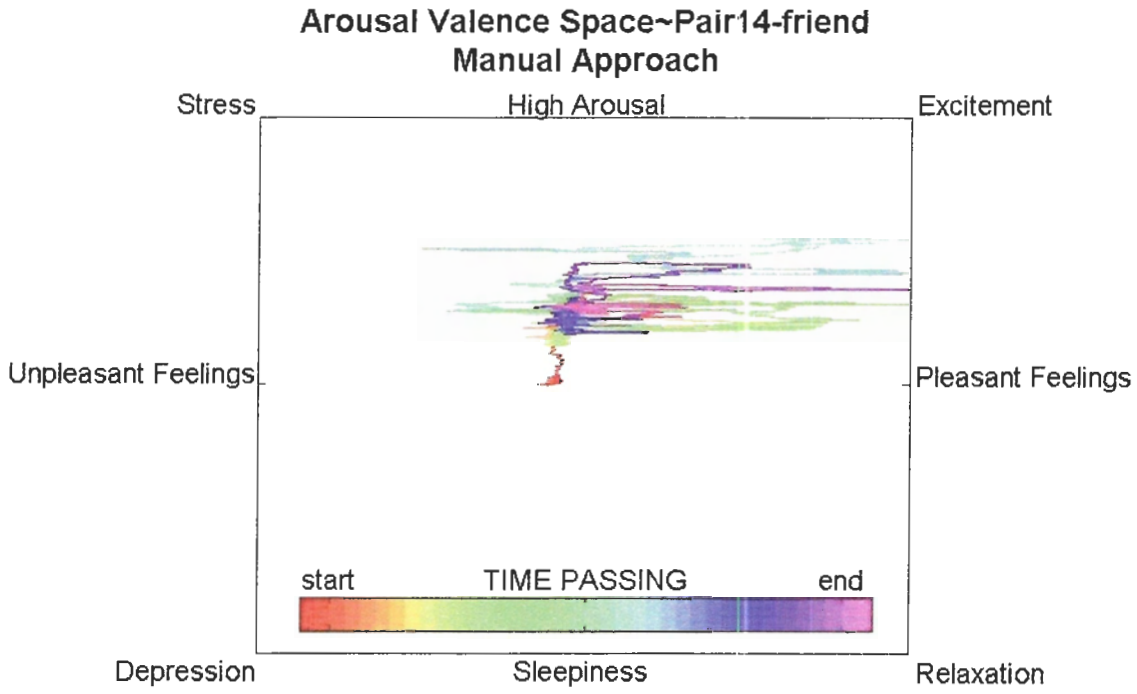

Arousal Valence Space~Pair5-friend
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

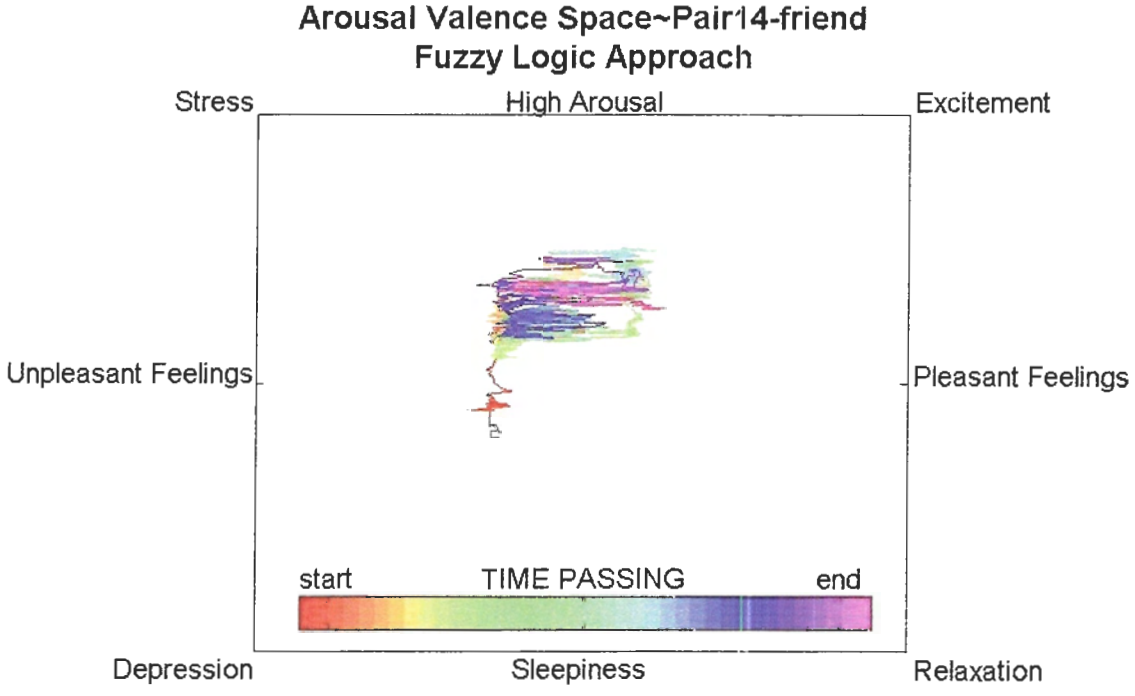

Arousal Valence Space~Pair5-stranger
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

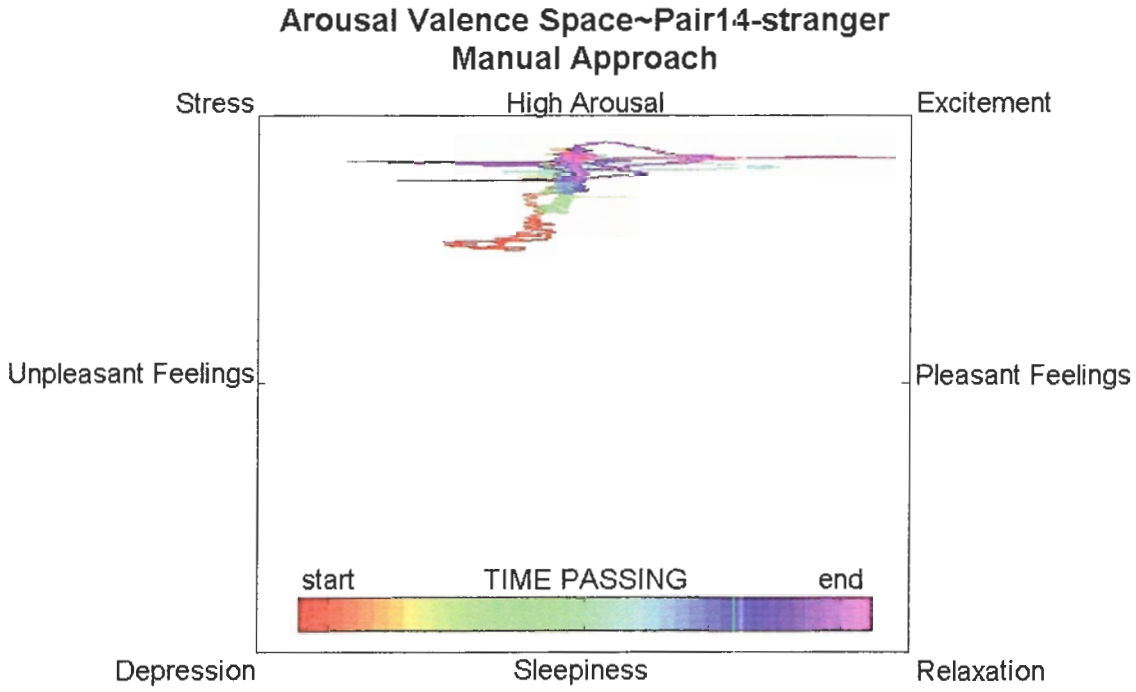

Arousal Valence Space~Pair5-stranger
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

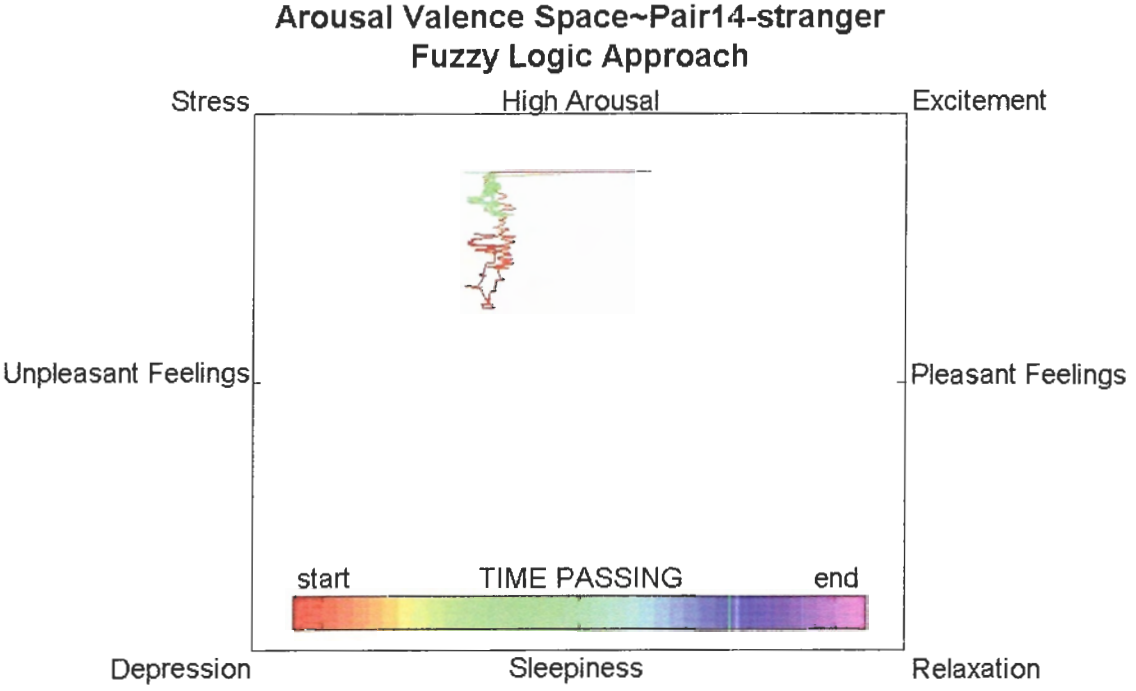

Arousal Valence Space~Pair14-computer
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

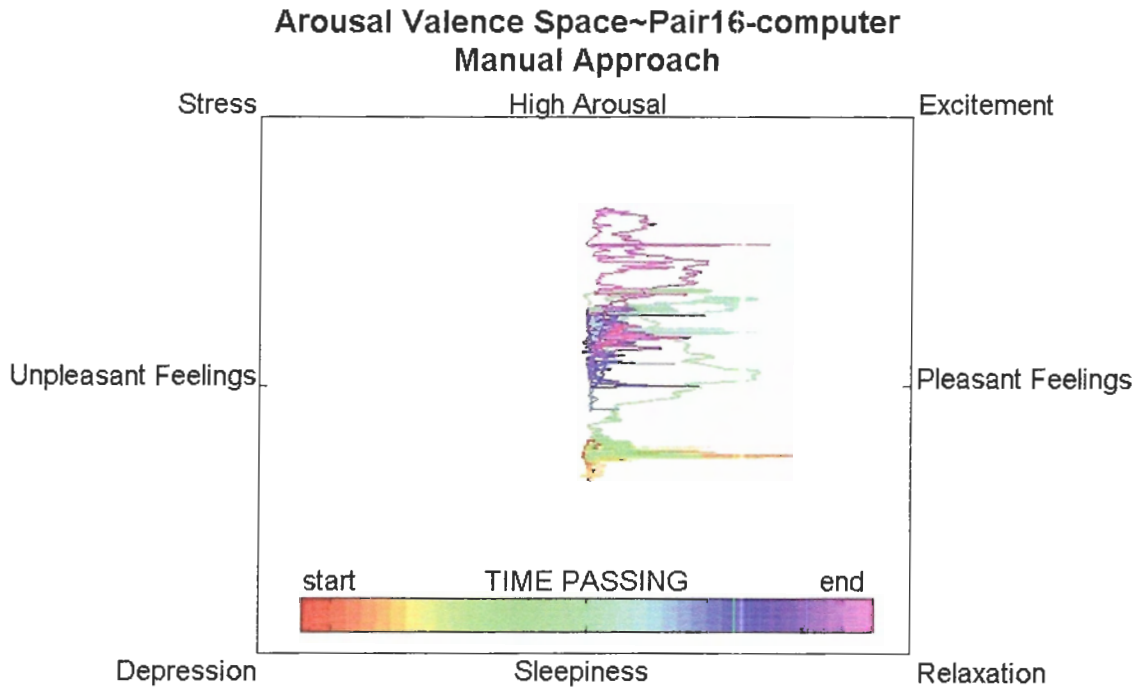

Arousal Valence Space~Pair14-computer
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

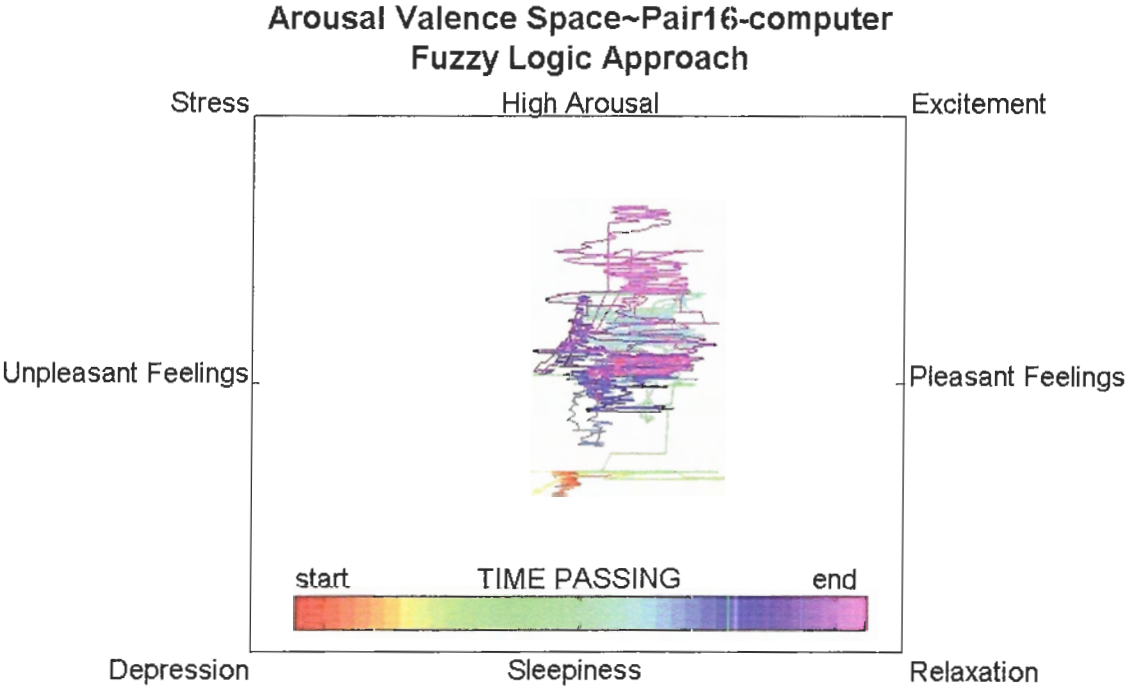

Arousal Valence Space~Pair14-friend
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

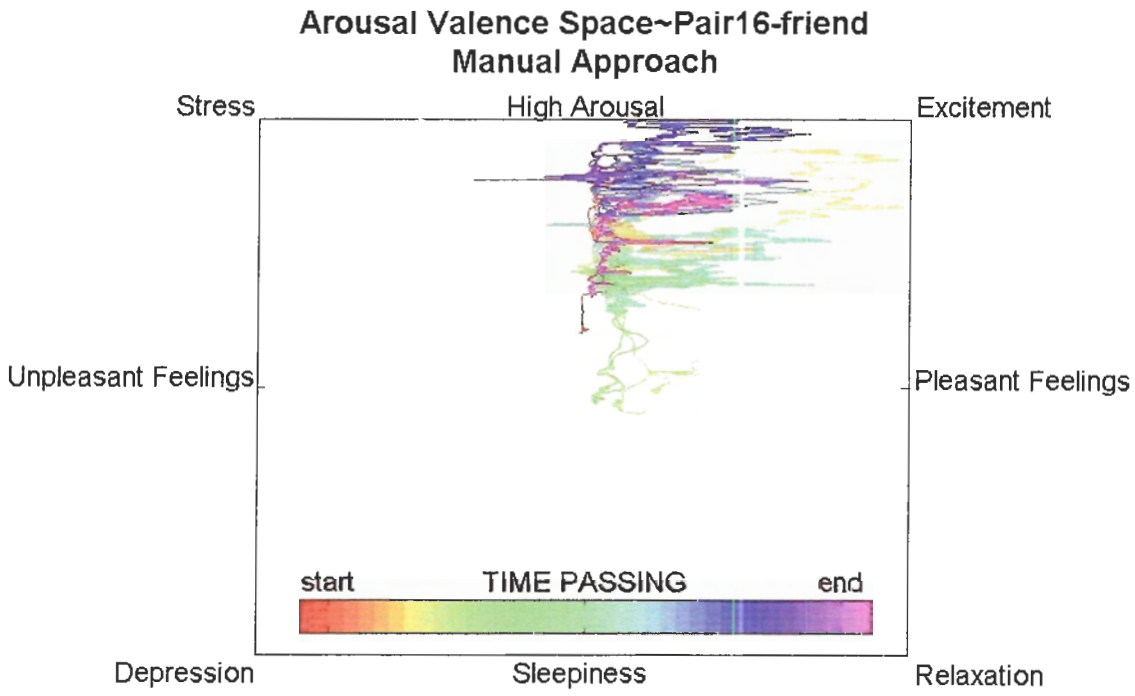

Arousal Valence Space~Pair14-friend
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

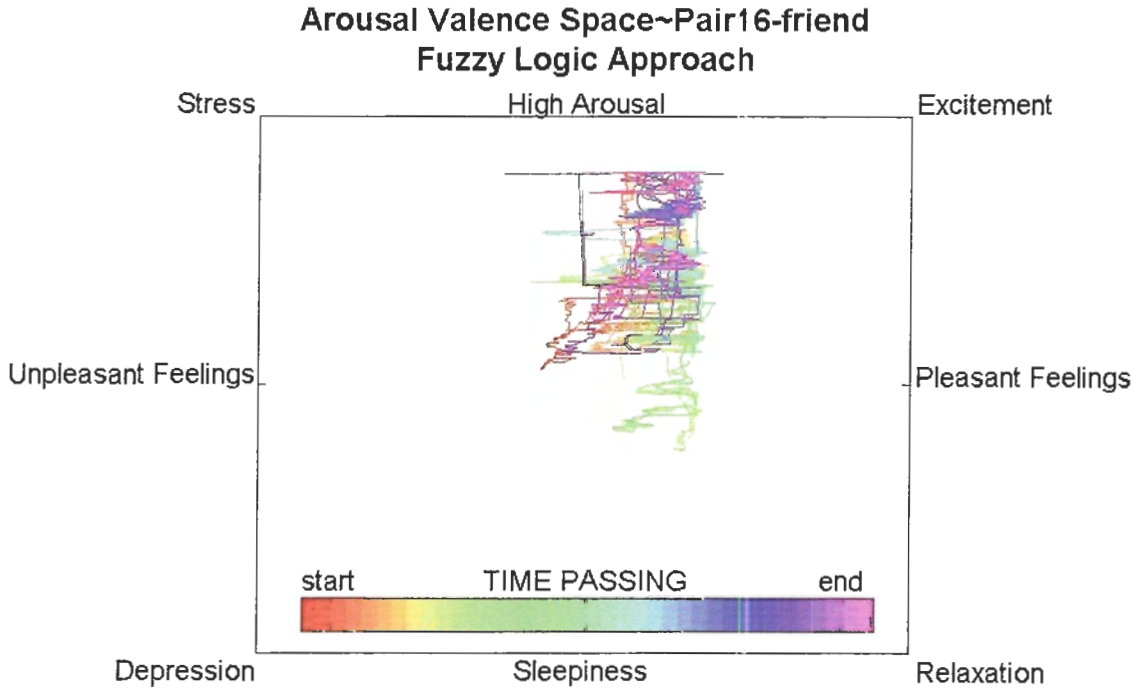

Arousal Valence Space~Pair14-stranger
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

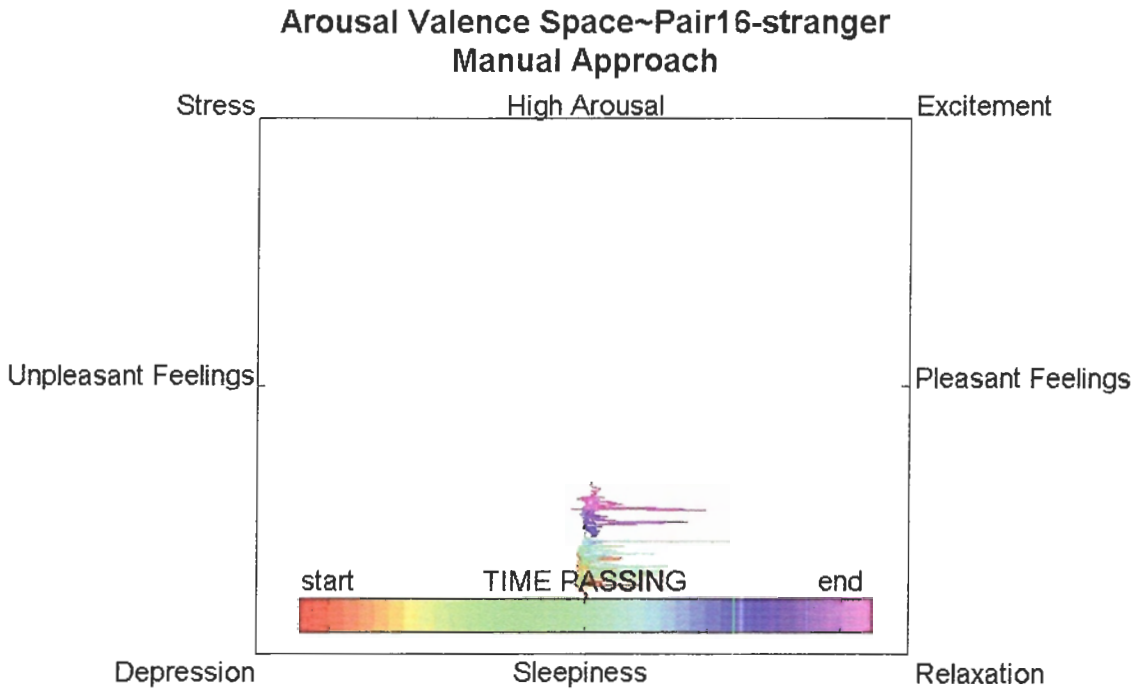

Arousal Valence Space~Pair14-stranger
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

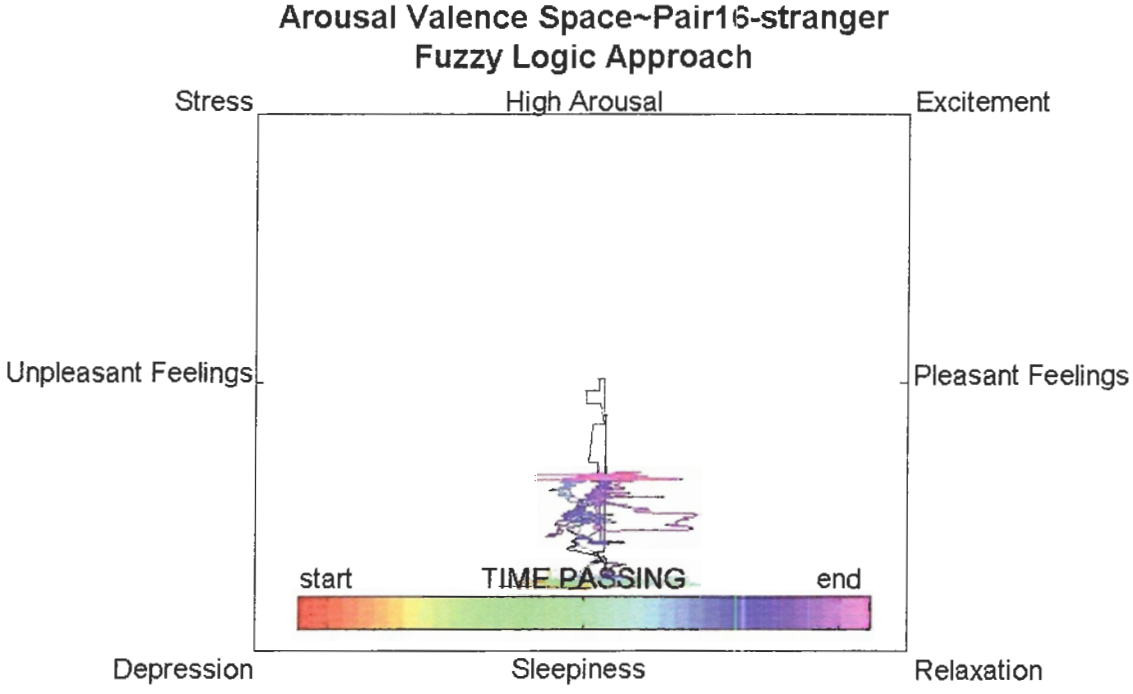

Arousal Valence Space~Pair16-computer
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

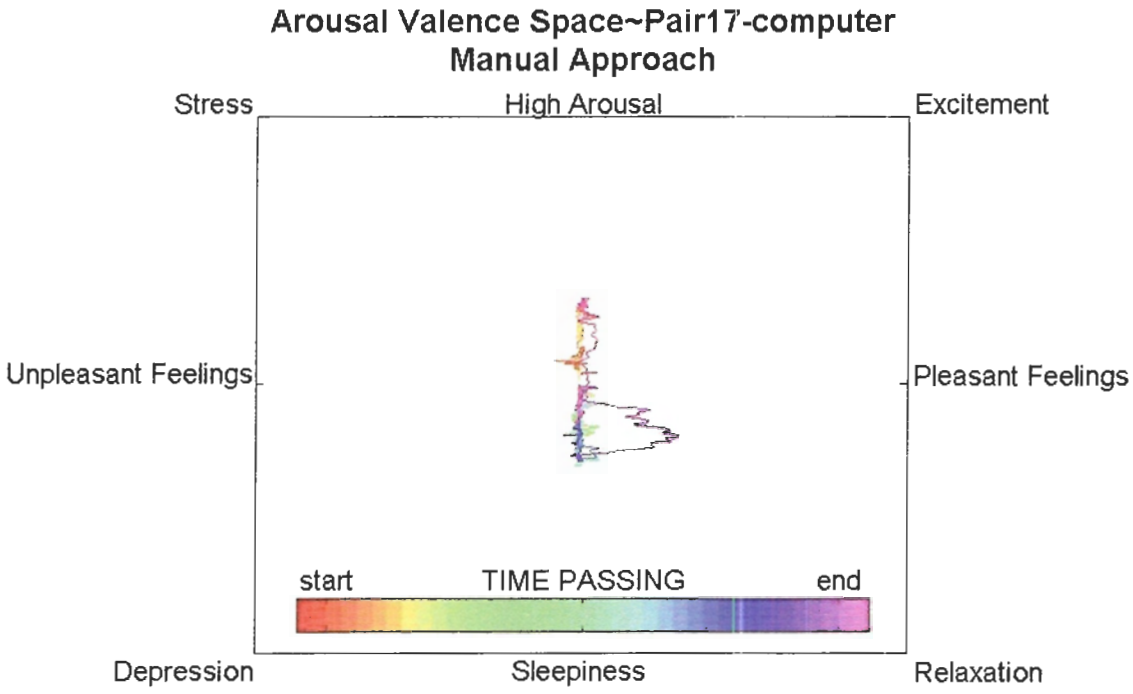

Arousal Valence Space~Pair16-computer
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

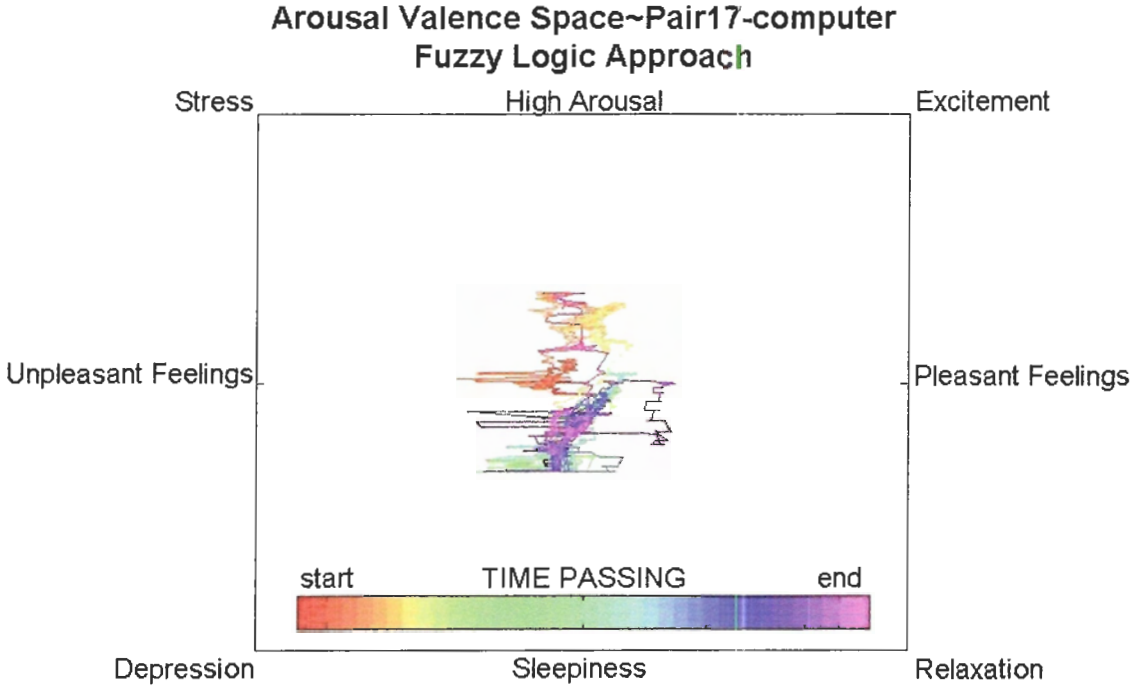

Arousal Valence Space~Pair16-friend
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

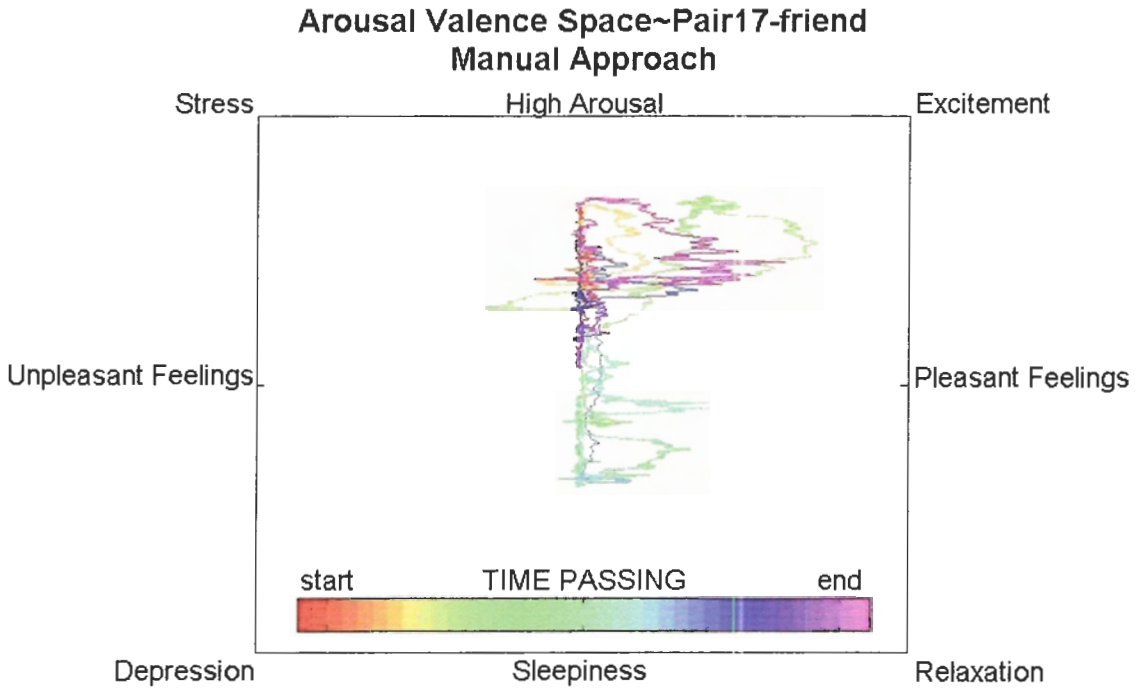

Arousal Valence Space~Pair16-friend
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

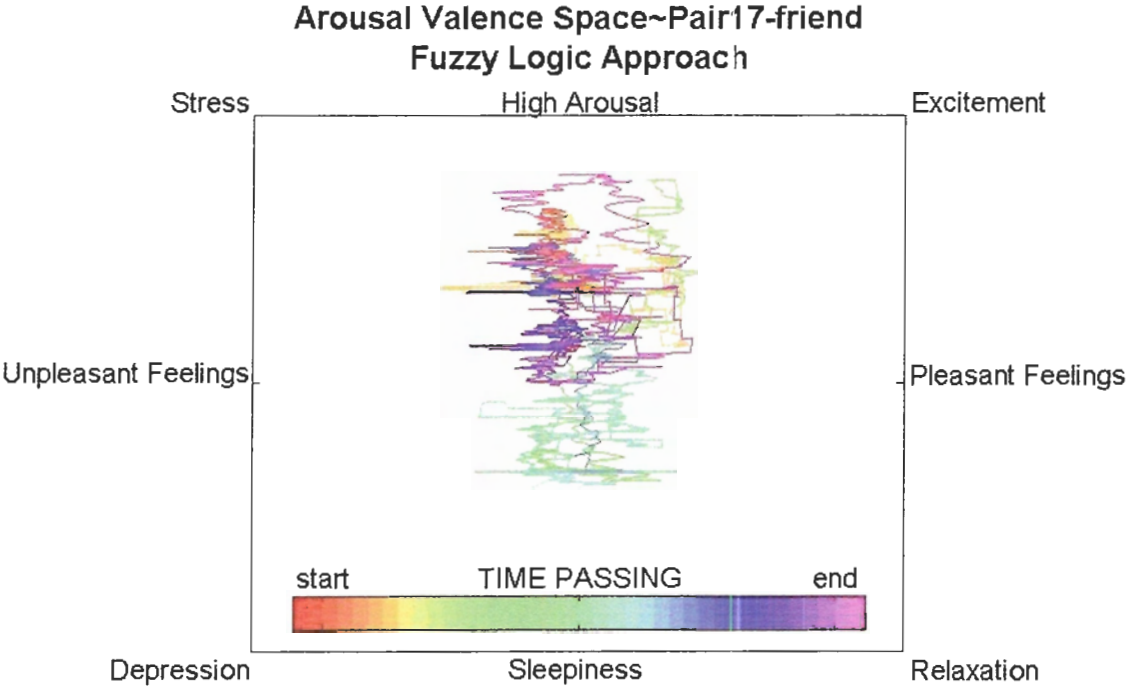

Arousal Valence Space~Pair16-stranger
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

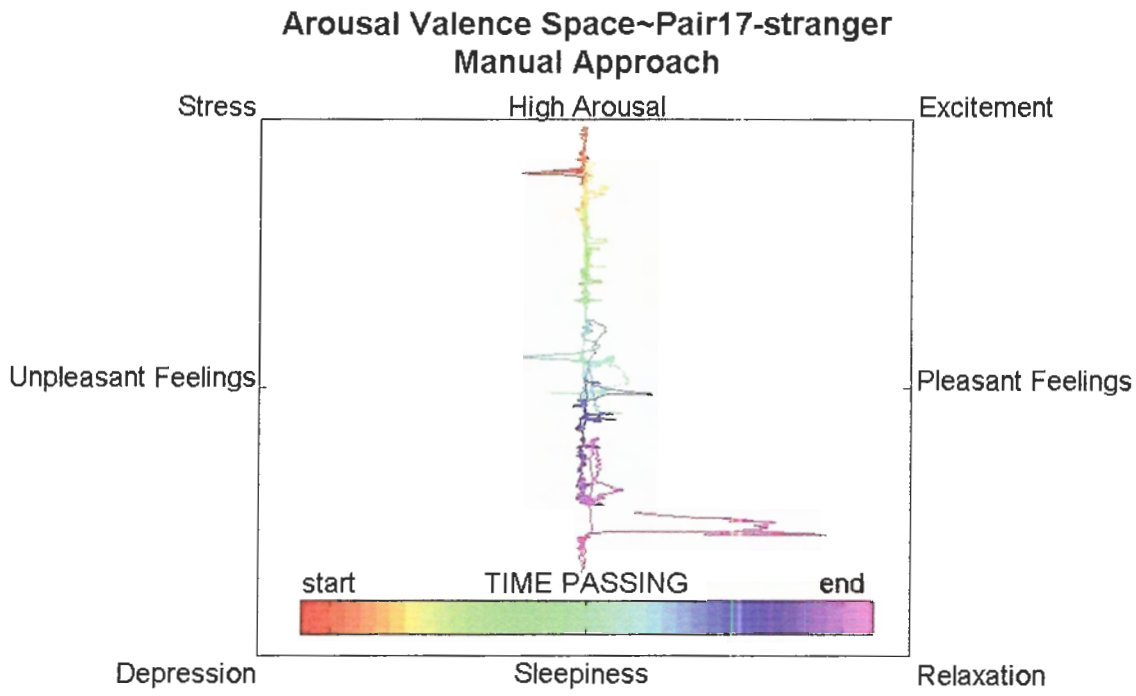

Arousal Valence Space~Pair16-stranger
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

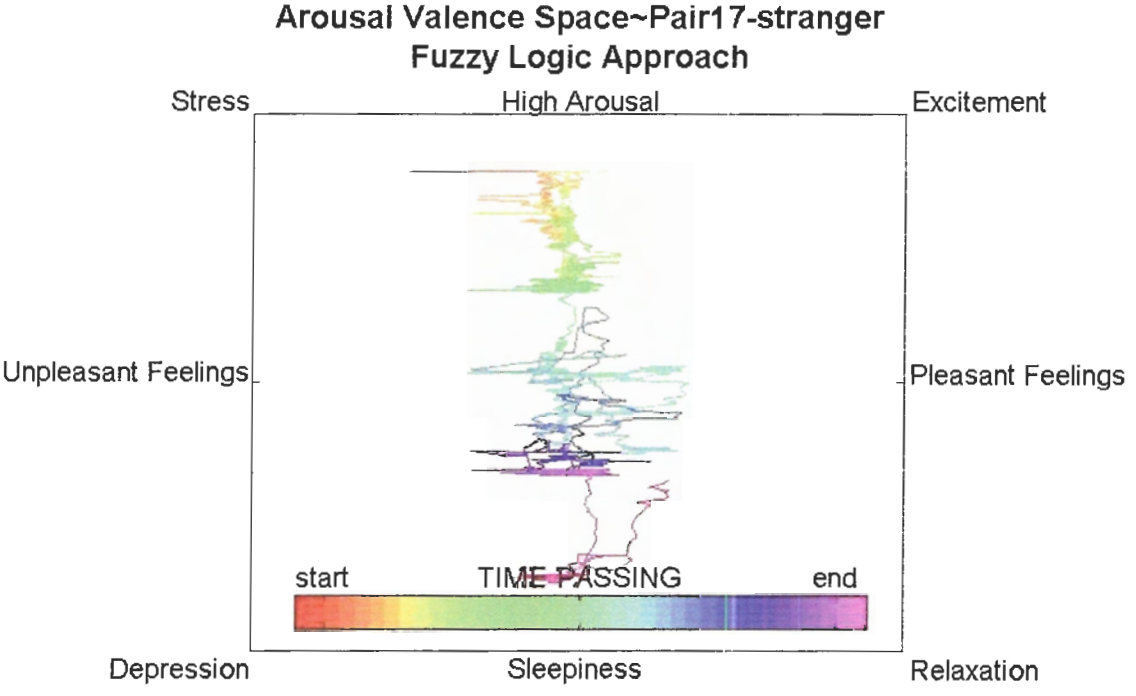

Arousal Valence Space~Pair17-computer
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

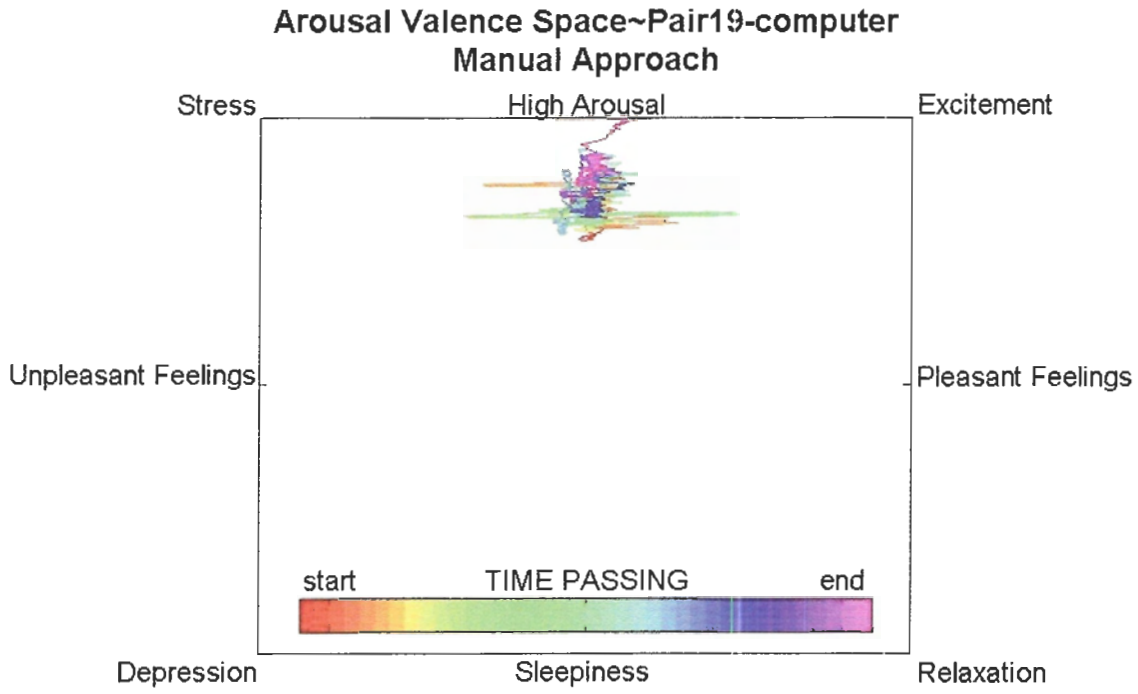

Arousal Valence Space~Pair17-computer
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

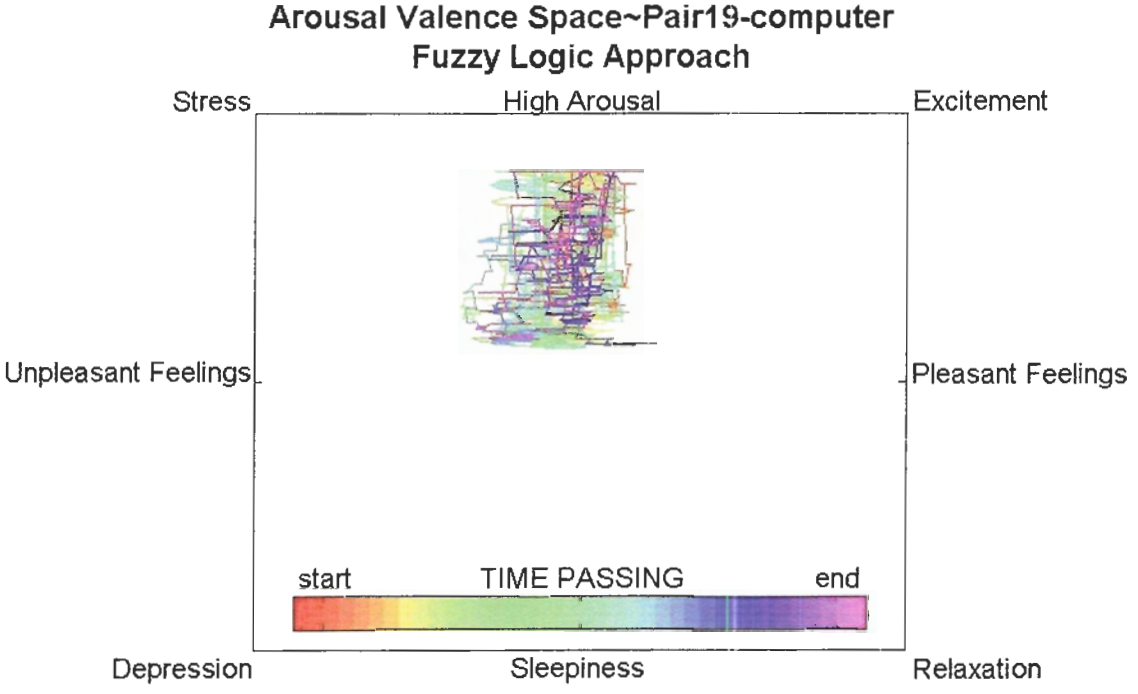

Arousal Valence Space~Pair17-friend
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

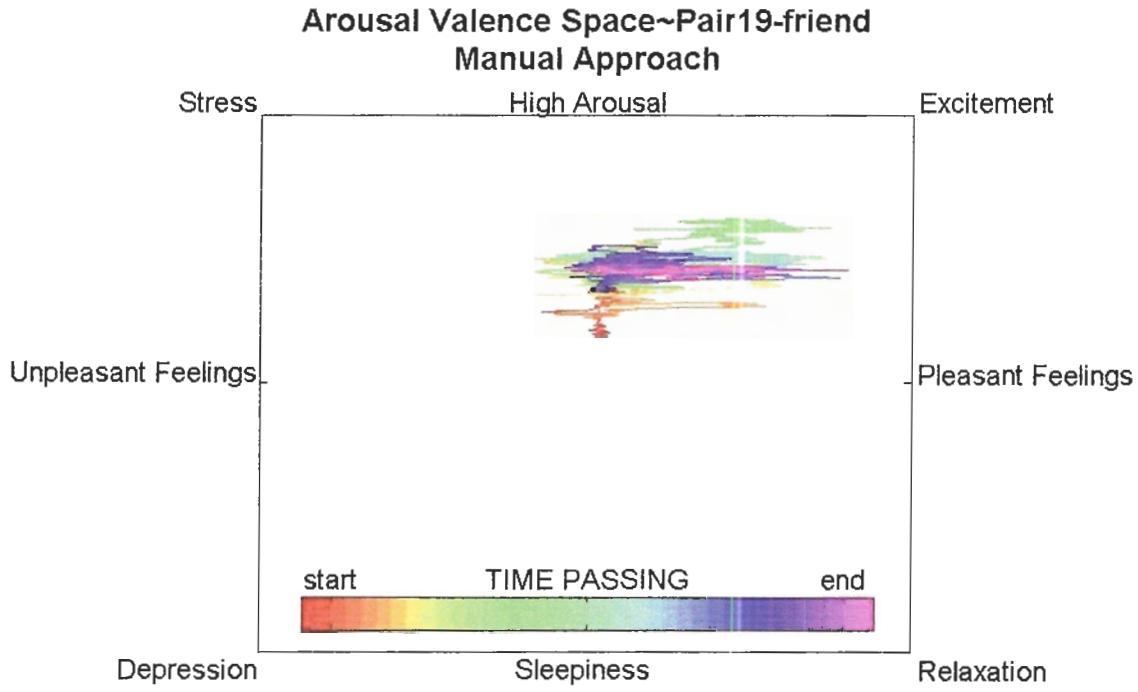

Arousal Valence Space~Pair17-friend
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

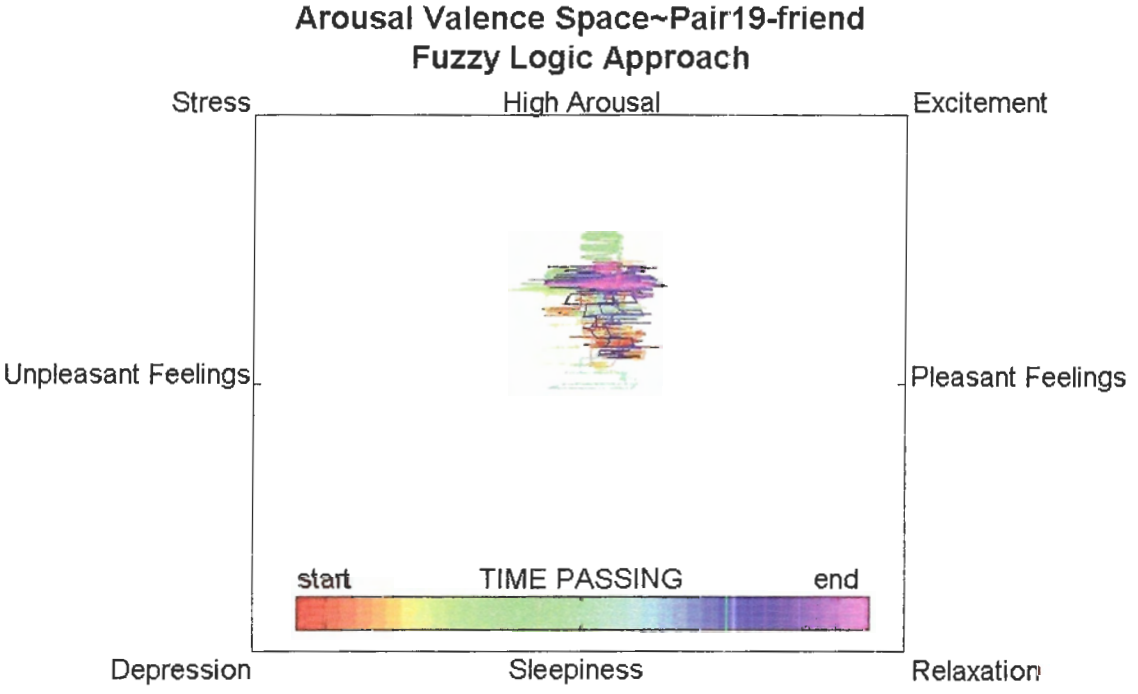

Arousal Valence Space~Pair17-stranger
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

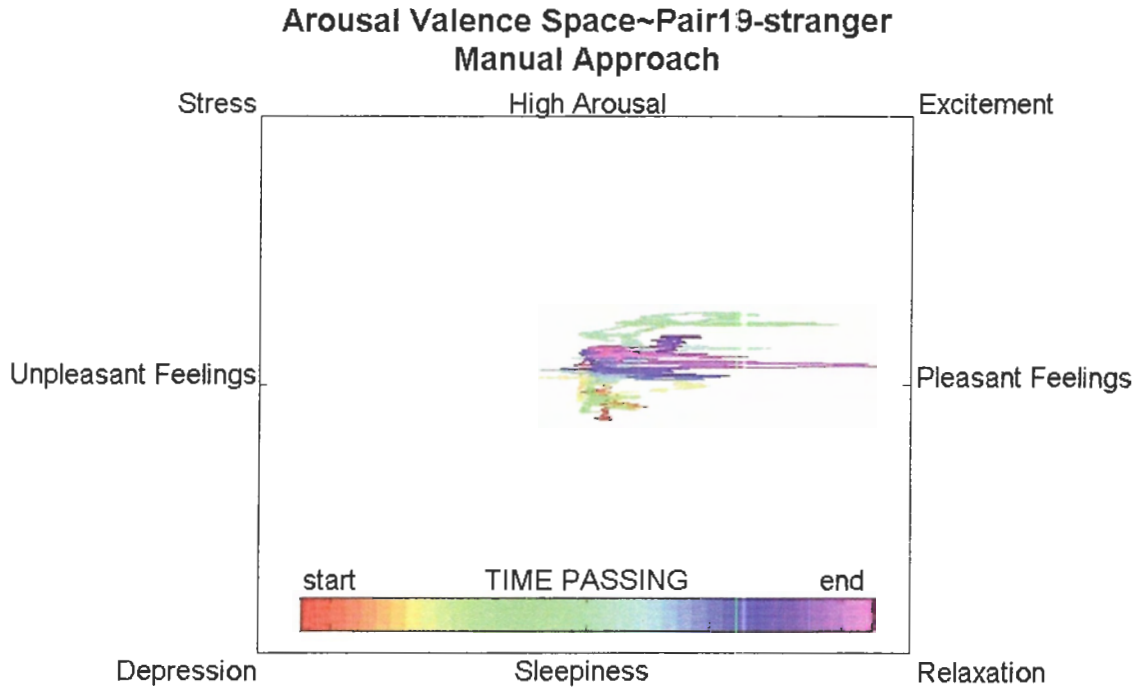

Arousal Valence Space~Pair17-stranger
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

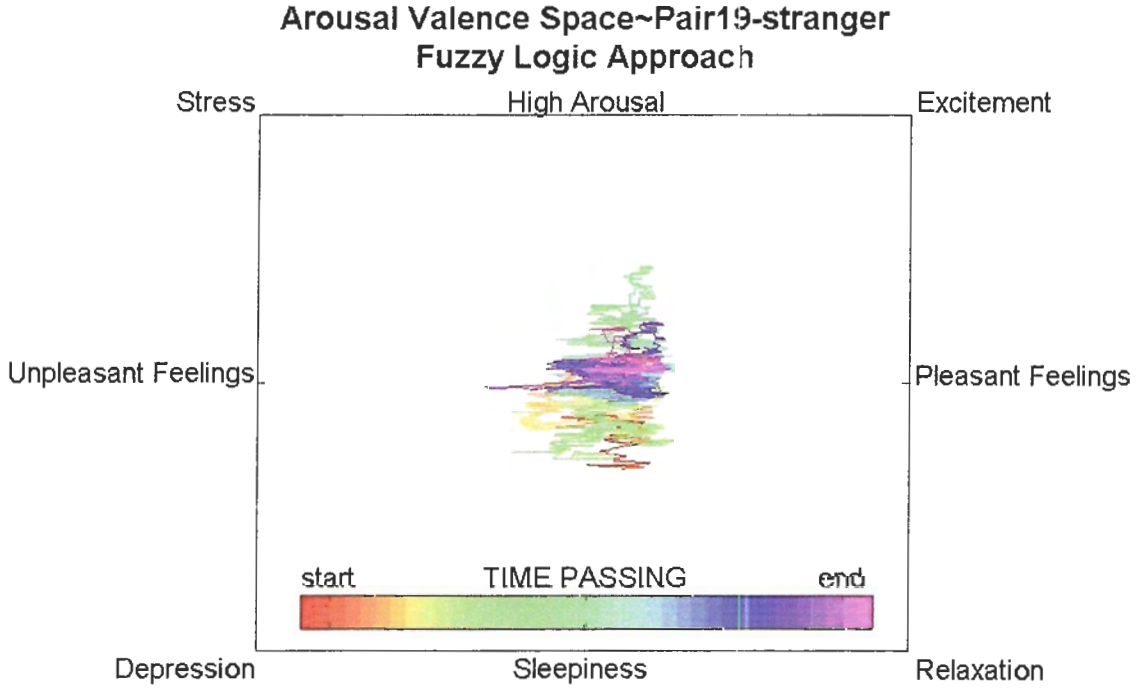

Arousal Valence Space~Pair19-computer
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

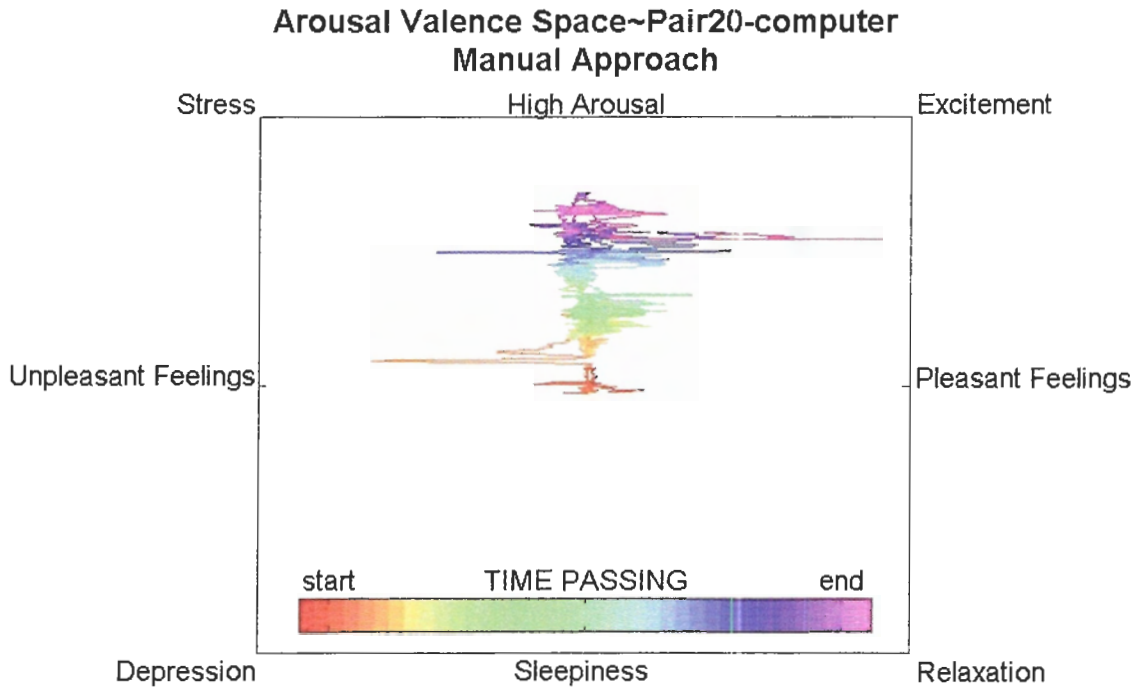

Arousal Valence Space~Pair19-computer
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

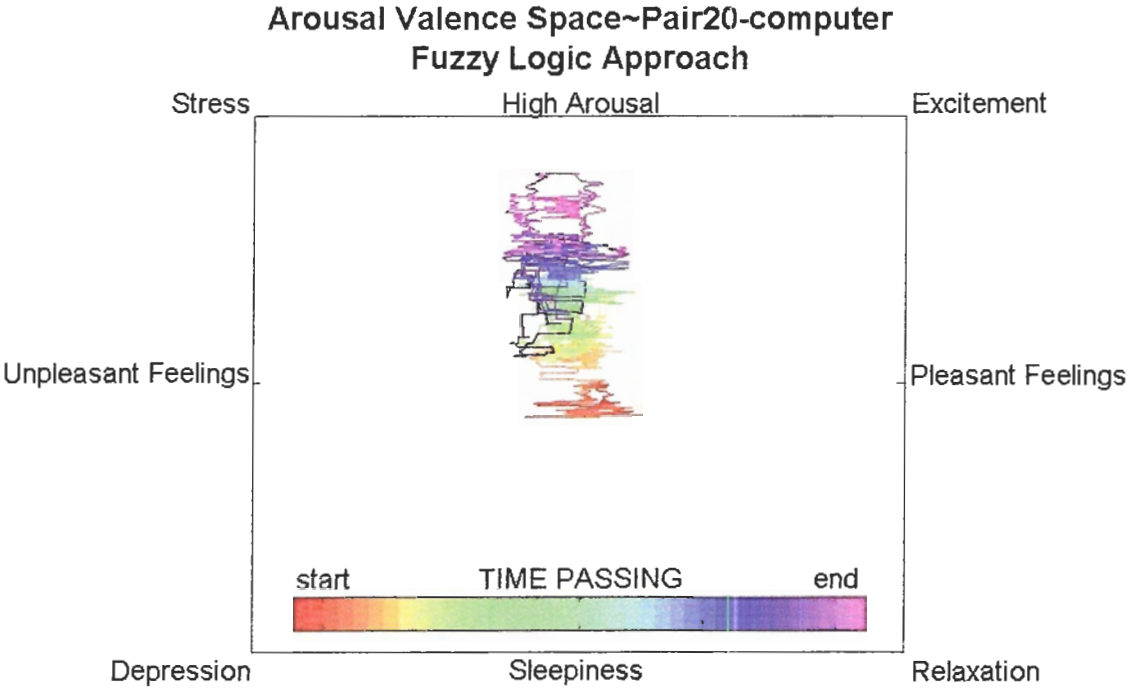

Arousal Valence Space~Pair19-friend
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

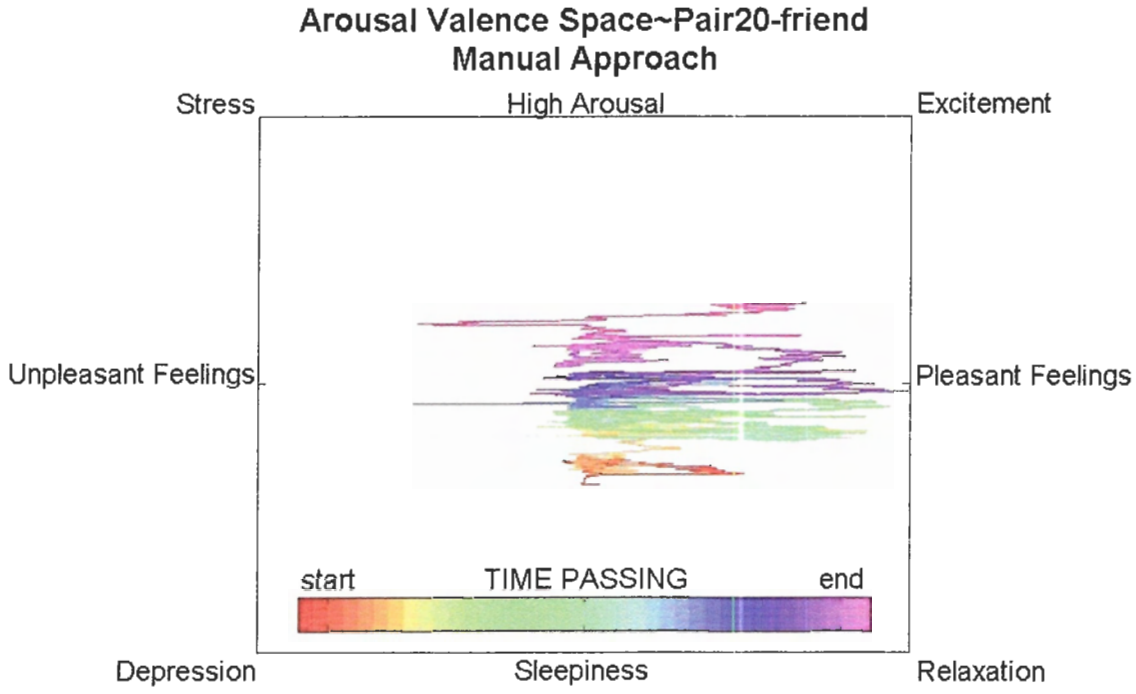

Arousal Valence Space~Pair19-friend
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

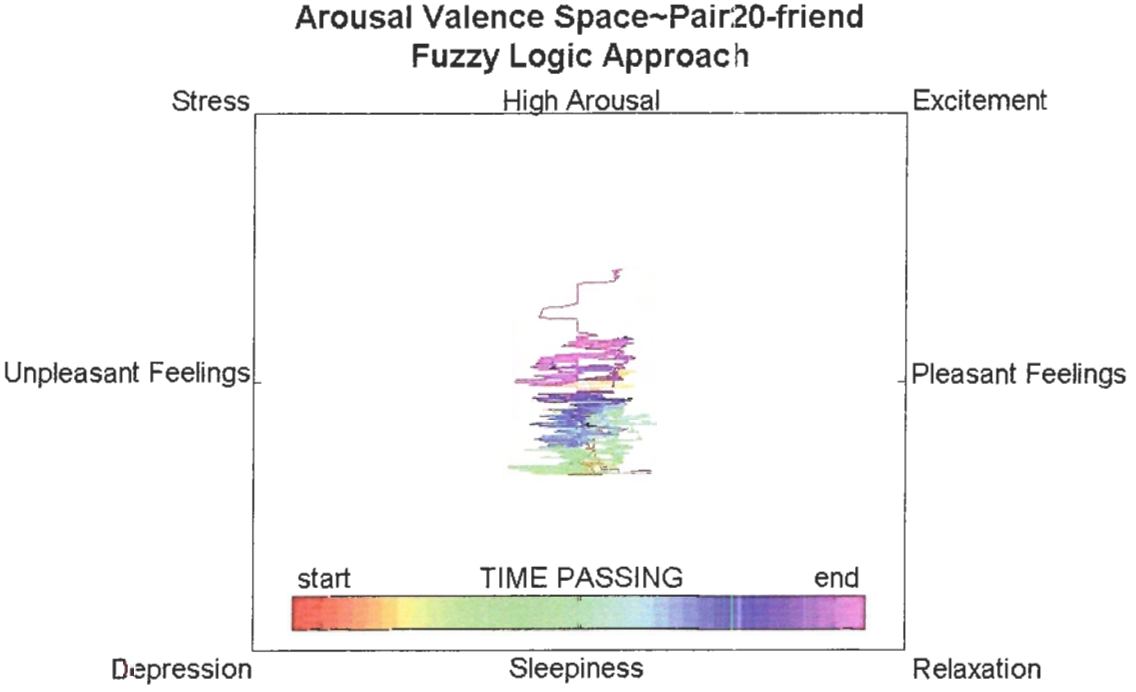

Arousal Valence Space~Pair19-stranger
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

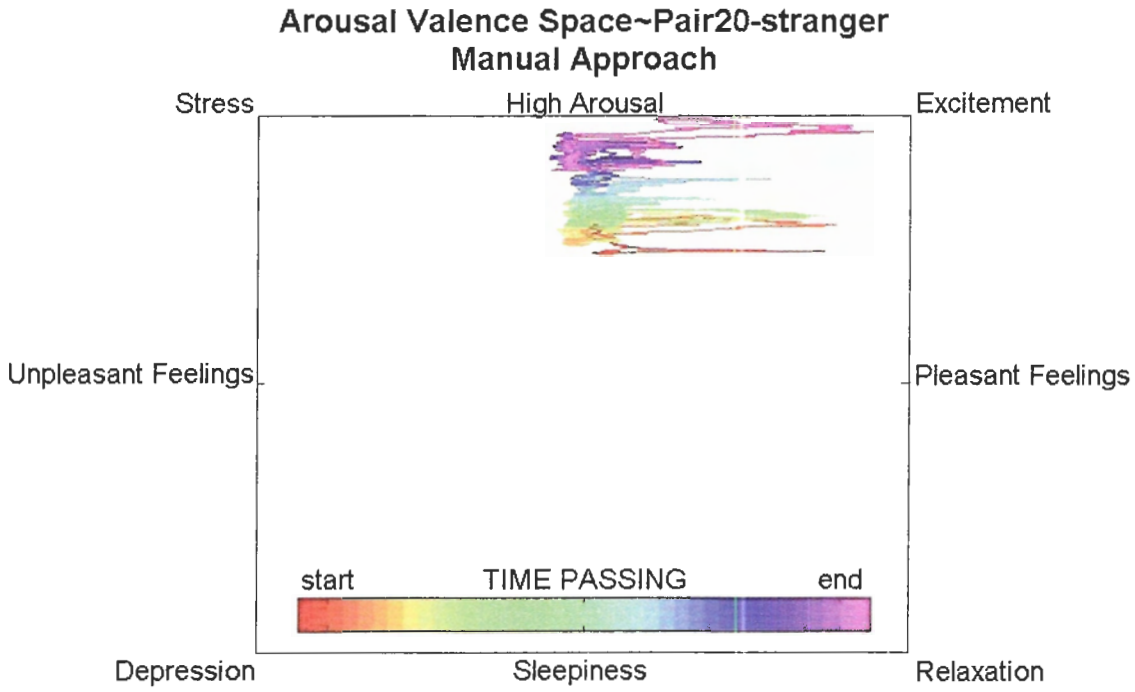

Arousal Valence Space~Pair19-stranger
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation

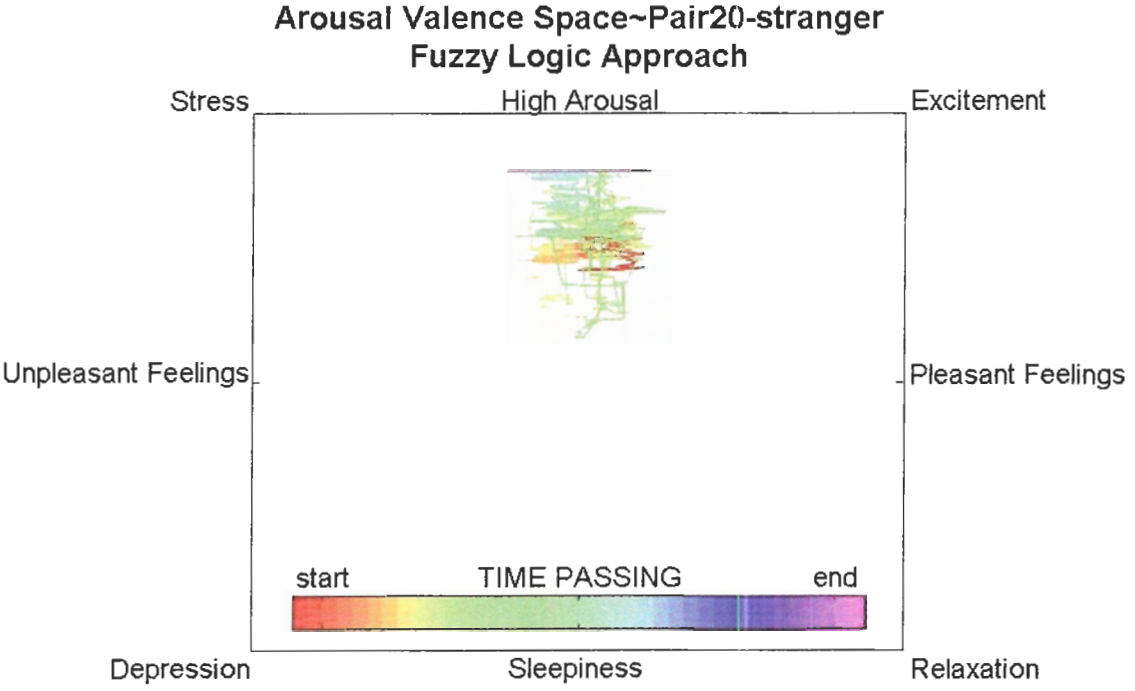

Arousal Valence Space~Pair20-computer
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation


Arousal Valence Space~Pair20-computer
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation


Arousal Valence Space~Pair20-friend
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation


Arousal Valence Space~Pair20-friend
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation


Arousal Valence Space~Pair20-stranger
Manual Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation


Arousal Valence Space~Pair20-stranger
Fuzzy Logic Approach
Stress
High Arousal
Excitement
Unpleasant Feelings
Pleasant Feelings
start
TIME PASSING
end
Depression
Sleepiness
Relaxation