Behaviour & Information Technology (in press)

Using psychophysiological techniques to measure user experience with entertainment technologies

REGAN L. MANDRYK*, KORI M. INKPEN[±] and THOMAS W. CALVERT*

*School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6 Canada, rlmandry@sfu.ca
[±]School of Computer Science, Dalhousie University, Halifax, NS B3H 1W5 Canada

Emerging technologies offer exciting new ways of using entertainment technology to create fantastic play experiences and foster interactions between players. Evaluating entertainment technology is challenging because success isn't defined in terms of productivity and performance, but in terms of enjoyment and interaction. Current subjective methods of evaluating entertainment technology aren't sufficiently robust. This paper describes two experiments designed to test the efficacy of physiological measures as evaluators of user experience with entertainment technologies. We found evidence that there is a different physiological response in the body when playing against a computer versus playing against a friend. These physiological results are mirrored in the subjective reports provided by the participants. In addition, we provide guidelines for collecting physiological data for user experience analysis, which were informed by our empirical investigations. This research provides an initial step towards using physiological responses to objectively evaluate a user's experience with entertainment technology.

1. Introduction

Emerging technologies in ubiquitous computing and ambient intelligence offer exciting new interface opportunities for co-located entertainment technology, as evidenced in a recent growth in the number of conference workshops and research articles devoted to this topic (Björk et al., 2001; Björk et al., 2002; Magerkurth et al., 2003). Our research team is interested in employing these new technologies to foster interactions between users in co-located, collaborative entertainment environments. We want technology to not only enable fun, compelling experiences, but also to enhance interaction and communication between players.

For example, we recently created a hybrid board-video game system to enhance player interaction (Mandryk et al., 2002). 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, do not 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. Our approach was to build a hybrid game system to leverage the advantages of both of these mediums, encouraging interaction between the players. We also created a collaborative game environment on handheld computers where players work together but individually access a shared game space, to enhance collaboration (Danesh et al., 2001; Mandryk et al., 2001). Players began with a limited set of genetic material for alien beings, and were encouraged to trade and breed their creatures to create a target creature. In order to visualize the potential outcome of breeding two creatures, we created a *What-If* feature. This feature semantically partitioned the data across multiple devices, encouraging the players to collaborate (Mandryk et al., 2001).

We created these environments with the goal of enhancing interaction between players and to create a compelling experience. Other researchers have used emerging technologies to create entertainment environments with the same goal in mind (Holmquist et al., 1999; Björk et al., 2001; Magerkurth et al., 2003). However, evaluating the success of these new interaction techniques and environments is an open research challenge.

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 (Norman, 2002). Although the study of human cognition has made significant progress in the last decade, the notion of emotion, which is equally important to design (Norman, 2002), is still not well understood, especially when the primary goals are to challenge and entertain the user. This approach presents a shift in focus from *usability* analysis to *user experience* analysis. Traditional objective measures used for productivity environments, such as time and accuracy, are not directly relevant to collaborative play.

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, but are more interested in whether our environment fosters interaction and communication between the players,

creates an engaging experience, and is fun. A successful interaction technique should provide seamless access to the game environment and be a source of fun in itself. Although traditional usability issues may still be relevant, they are subordinate to the actual playing experience as defined by challenge, engagement, and fun.

Once a definition of success has been determined, we need to resolve how to measure the chosen variables. Unlike performance measures, the measures of success for collaborative entertainment technologies are more elusive. The current research problem lies in what metrics to use to measure engagement, interaction, and fun. These metrics will likely be of great interest to companies whose business success depends on developing successful games.

1.1. Evaluation of entertainment technologies

Current methods of evaluating collaborative entertainment include both subjective and objective techniques. The most common methods include subjective self-reports through questionnaires, interviews, and think-aloud protocols, and objective reports through observational video analysis.

Subjective reporting through questionnaires and interviews is generalizable, convenient, amenable to rapid statistics and easy to administer. Some drawbacks of questionnaires and surveys are that they are not conducive to finding complex patterns, can invade privacy, and subject responses may not correspond to the actual experience (Marshall & Rossman, 1999). Knowing that their answers are being recorded, participants will sometimes answer what they think you want to hear, without realizing it. Subjective ratings are cognitively mediated, and may not accurately reflect what is occurring (Wilson & Sasse, 2000b).

Think-aloud techniques (Nielsen, 1992), 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. To avoid disrupting the player during the game, we have previously 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.

Using video to code gestures, body language, and verbalizations is a rich source of data. Analysis techniques of observational data include conversation analysis, verbal and non-verbal protocol analysis, cognitive task analysis, and discourse analysis (Fisher & Sanderson, 1996). 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. Researchers must be careful to acknowledge their biases, address inter-rater reliability, and not read inferences where none are present (Marshall & Rossman, 1999). There is an enormous time commitment associated with observational analysis. The analysis time to data sequence time ratio (AT:ST) typically ranges from 5:1 to 100:1 (Fisher & Sanderson, 1996). Consequently, many researchers rely on subjective data for user preference, rather than objective observational analysis.

We have previously used both subjective reports and video coding as methods of evaluating our novel entertainment technologies, although there has been no control environment with which to make comparisons (Mandryk et al., 2001; Mandryk et al., 2002; Scott et al., 2003).

Researchers in Human Factors have used physiological measures as indicators of mental effort and stress (Vicente et al., 1987; Wilson, 2001). Psychologists use physiological measures as unique identifiers of human emotions such as anger, grief, and sadness (Ekman et al., 1983). However, physiological data have not been employed to identify user experience states such as engagement and fun. Our research aims to uncover whether there are links and correlations between a player's physiological state, events occurring during the entertainment experience, and subjected reported experience. These correlations would enable the testing and evaluation of novel collaborative entertainment technologies, such as the systems previously presented, in terms of enhancing interaction and increasing engagement and fun. Based on previous research on the use of psychophysiological techniques, we believe that directly capturing and measuring autonomic nervous system (ANS) activity will provide researchers and developers of technological systems with access to the 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.

1.2. Overview of research

The goal of our research was to test the efficacy of physiological measures for use in evaluating player experience with collaborative entertainment technologies. We have two 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.*

This paper describes two experiments that we designed to test the two main conjectures. We record users' physiological, verbal and facial reactions to game technology, and apply post-processing techniques to correlate an individual's physiological data with their subjective reported experience and events in the game. Our ultimate goal is to create a methodology for the objective evaluation of collaborative entertainment technology, as rigorous as current methods for productivity systems.

To provide an introduction for readers unfamiliar with physiological measures, we briefly introduce the physiological measures used, describe how these measures are collected, and explain their inferred meaning. We then present two experiments designed to investigate the applicability of physiological measures as indicators of human experience with entertainment technologies. The first experiment manipulated game difficulty and is described in section 4. Throughout the description of Experiment One, we provide information on how to approach the collection and analysis of physiological signals. Based on the lessons we learned, and the results from Experiment One, we

conducted Experiment Two, which is described in section 5. Finally, we discuss our plans for future work, and conclude with a summary of the results, and a description of caveats for conducting this type of research.

2. Physiology and Emotions

In our research, physiological data were gathered using the Procomp Infiniti hardware 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), 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. The measures we used will each be described briefly including reference to how they have previously been used in technical domains.

2.1. Galvanic skin response

GSR is a measure of the conductivity of the skin. There are specific sweat glands (eccrine glands) that cause this conductivity to change and result in the GSR . Located in the palms of the hands and soles of the feet, these sweat glands respond to psychological stimulation rather than simply to temperature changes in the body (Stern et al., 2001). For example, many people have cold clammy hands when they are nervous. In fact, subjects do not have to even be sweating to see differences in skin conductance in the palms of the hands or soles of the feet 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 reach the surface of the skin (Stern et al., 2001).

Galvanic skin response is a linear correlate to arousal (Lang, 1995) and reflects both emotional responses as well as cognitive activity (Boucsein, 1992). GSR has been used extensively as an indicator of experience in both non-technical domains (see (Boucsein, 1992) for a comprehensive review), and technical domains (Wilson & Sasse, 2000a; Wilson, 2001; Ward et al., 2002; Ward & Marsden, 2003).

We measured GSR using surface electrodes sewn in Velcro™ straps that were placed around two fingers on the same hand. Previous testing of numerous electrode placements was conducted to ensure that there was no interference from movements made when manipulating the game controller. We found no differences between responses from pre-gelled electrodes on the feet and responses from the finger clips we employed.

2.2. Cardiovascular measures

The cardiovascular system includes the organs that regulate blood flow through the body. Measures of cardiovascular activity include HR, interbeat interval (IBI), heart rate variability (HRV), blood pressure (BP), and BVP. EKG measures electrical activity of the heart. HR, HRV, and respiratory sinus arrhythmia (RSA) can all be gathered from EKG.

HR reflects emotional activity. It has been used to differentiate between positive and negative emotions with further differentiation made possible with finger temperature (Winton et al., 1984; Papillo & Shapiro, 1990). HRV refers to the oscillation of the interval between consecutive heartbeats. When subjects are under stress, HRV is suppressed and when they are relaxed, HRV emerges. Similarly, HRV decreases with mental effort, but if the mental effort needed for a task increases beyond the capacity of working memory, HRV will increase (Rowe et al., 1998).

Although there is a standard medical configuration for placement of electrodes, two electrodes placed fairly far apart will produce an EKG signal (Stern et al., 2001). We placed three pre-gelled surface electrodes in the standard configuration of two electrodes on the chest and one electrode on the abdomen.

2.3. Respiratory measures

Respiration can be measured as the rate or volume at which an individual exchanges air in their lungs. Rate of respiration (RespRate) and depth of breath (RespAmp) are the most common measures of respiration.

Emotional arousal increases respiration rate while rest and relaxation decreases respiration rate (Stern et al., 2001). Although respiration rate generally decreases with relaxation, startle events and tense situations may result in momentary respiration cessation. Negative emotions cause irregularity in the respiration pattern (Stern et al., 2001). Because respiration is closely linked to cardiac function, a deep breath can affect cardiac measures.

Respiration is most accurately measured by gas exchange in the lungs, but the sensor technology inhibits talking and moving (Stern et al., 2001). Instead, chest cavity expansion can be used to capture breathing activity using either a Hall effect sensor, strain gauge, or a stretch sensor (Stern et al., 2001). We used a stretch sensor sewn into a Velcro™ strap, positioned around the thorax.

2.4. Electromyography

Electromyography (EMG) measures muscle activity by detecting surface voltages that occur when a muscle is contracted (Stern et al., 2001). In isometric conditions (no movement) EMG is closely correlated with muscle tension (Stern et al., 2001), 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 (Cacioppo et al., 2000). On the face, EMG has been used to distinguish between positive and negative emotions. EMG activity over the brow (frown muscle) region is lower and EMG activity over the cheek (smile muscle) is higher when emotions are mildly positive, as opposed to mildly negative (Cacioppo et al., 2000).

We used surface electrodes to detect EMG on the jaw, indicative of tension. 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.

2.5. Identifying Emotions

There has been a long history of researchers using physiological data to try to identify emotional states. William James first speculated that patterns of physiological response could be used to recognize emotion (Cacioppo & Tassinary, 1990), and although this viewpoint is too simplistic, recent evidence suggests that physiological data sources can differentiate among some emotions (Levenson, 1992). There are varying opinions on whether emotions can be classified into discrete, specific emotions (Ekman, 1999), or whether emotions exist along multiple axes in space (Russell et al., 1989; Lang, 1995). Both theoretical perspectives have seen limited success in using physiological data to identify emotional states, see (Cacioppo et al., 2000) for an overview. In addition to the difficulties in classifying emotions, when using physiological data sources there are methodological issues that must be addressed (Picard, 1997), and theoretical limitations to inferring significance (Cacioppo & Tassinary, 1990). Discussing these issues are beyond the scope of this paper.

3. Related Literature on Using Physiology as a Metric of Evaluation

Although there is no previous research on using physiology as an indicator of fun, or engagement with entertainment technology, or as an indicator of collaborative interaction, it has been used in other domains as a metric of evaluation.

The field of human factors has been concerned with optimizing the relationship between humans and their technological systems. The quality of a 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 is formed from the triangulation of three fundamental dimensions: overt behaviour, physiology, and subjective experience (Wastell & Newman, 1996).

Wastell and Newman (Wastell & Newman, 1996) used the physiological measures of blood pressure (systolic and diastolic) 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 dispatcher 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.

Wilson and Sasse (Wilson & Sasse, 2000a, b; Wilson, 2001) used physiological measures to evaluate subject responses to audio and video degradations in videoconferencing software. The authors suggest that subjective ratings of user satisfaction and objective measures of task performance be augmented with physiological measures of user cost (Wilson, 2001). 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 (Wilson & Sasse, 2000a), even though most 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 (Wilson & Sasse, 2000b), but these results weren't always consistent with subjective responses. These discrepancies between physiological and subjective assessment support the argument for a three-tiered approach.

Ward et al. (Ward et al., 2002; Ward & Marsden, 2003) collected GSR, BVP, and HR while subjects attempted to answer questions by navigating through both well and ill designed web pages. No significant differences were found between users of the two types of web pages, which is not surprising considering the large individual differences associated with physiological data. 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 GSR and level pulse rate).

These studies collected both subjective measures and physiological data, however, did not try to correlate the two data sources using normalized measures. Using a hovercraft simulator, Vicente et al. (Vicente et al., 1987) normalized heart rate variability (HRV) to a ratio between zero and one. They determined that normalized HRV data significantly correlated to subjective ratings of effort, but not workload or task difficulty. In the domain of HCI, a few other researchers have also used HRV as an indicator of mental effort (Rowe et al., 1998; Rani et al., 2002).

Partala and Surakka (Partala & Surakka, 2004) and Scheirer et al. (Scheirer et al., 2002) both used pre-programmed mouse delays to intentionally frustrate a computer user. Partala and Surakka measured EMG activity on the face in response to positive, negative, or no audio intervention, while Scheirer et al. applied Hidden Markov Models (HMMs) to GSR and BVP data to detect states of frustration.

In the domain of entertainment technology, Sykes and Brown (Sykes & Brown, 2003) measured the pressure that gamers exerted on the gamepad controls while participants played Space Invaders. They found that the players exerted more pressure in the difficult condition than in the easy or medium conditions. They did not correlate the pressure data with any type of subjective report.

Although very little research has been conducted in the entertainment domain, results from the few studies in HCI and the more plentiful studies in the field of human factors are encouraging. The studies presented in this section each reveal how different physiological measures were successfully used in different work-related domains, however, the emerging nature of this technique means that there has been no standardization of task, domain, or measures. As such, comparison across studies is difficult. Building a corpus of knowledge surrounding the use of physiological measures in HCI evaluation is occurring, albeit slowly. There is still a need for researchers from the fields of psychology, kinesiology, HCI, machine learning, and pattern recognition, who are interested in physiological techniques for

HCI evaluation, to create a research community in order to advance the fledgling field.

4. Experiment One

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.*

4.1. Participants and setting

Eight male participants were recruited from computer science and engineering students at Simon Fraser University to participate in the experiment. One participant did not complete the experiment, so we present data for seven participants aged 20 to 26. All participants filled out a background questionnaire, which 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. 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.

All participants were frequent computer users. When asked to rate how often they used computers on a five-point scale, all seven subjects used them every day (corresponding to five). 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. When asked how much they liked different game genres, role-playing was the favorite, followed by strategy and action games.

Participants played the game at four conditions of difficulty: beginner, easy, medium, and difficult. To balance the order of presentation of the difficulty conditions, we used a reversed Latin Square design. Participants played NHL 2003 by EA Sports™ in all conditions (see Figure 1 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. 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.

Figure 1: Screen shot of NHL 2003 by EA Sports™.

Each play condition consisted of one five-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.

The experiment was conducted at the New Media Innovation Centre, in Vancouver, Canada. NHL 2003™ was played on a Sony PS2™, and viewed on a 36" television. Cameras captured the 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 2). The video was used to determine the times of the rest period and four difficulty conditions, and for qualitative analysis, but not for quantitative video annotation.

Upon arriving, participants signed a consent form, after which they were fitted with the physiological sensors. The participants then rested for five minutes, after which they played the game in their first difficulty level. After each difficulty condition, the primary experimenter interviewed the participants, asking them to rate the challenge, frustration, boredom, and fun of each condition on a scale of one (low) to five (high). Explanation of their answers was encouraged. After completing the experiment, the same experimenter interviewed the

participants again, asking them to rank the four difficulty conditions in terms of challenge, excitement, and fun. Again, they were encouraged to explain their answers.

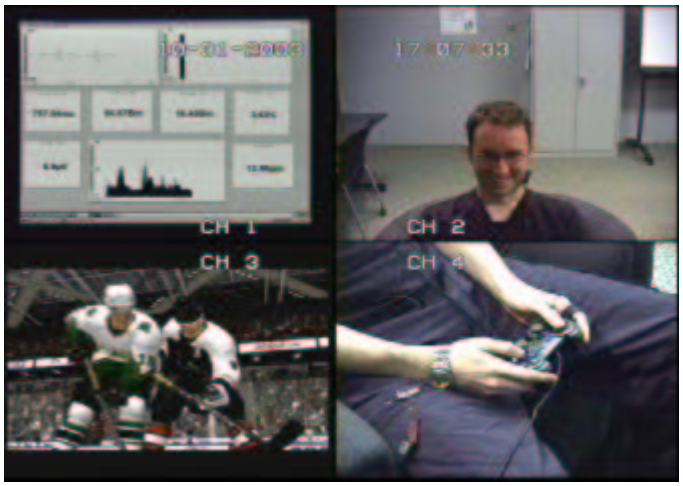

Figure 2: Quadrant display: screen capture of biometrics, video of player's face, video of controller, and screen capture of game.

4.2. Data analyses

The subjective data from both the condition questionnaires and the post experiment questionnaires were collected into a database, and analysed 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, RespRate, and RespAmp were computed at 4 Hz. Physiological data for the 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 only one beat has been counted in two sampling intervals. We inspected the HR data and corrected these erroneous samples. For each condition and the 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 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.

4.3. Results and discussion

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

4.3.1. Subjective responses

Participants rated the boredom, challenge, frustration, and fun on a five-point scale after playing in each of the conditions. The mean results are shown in Table 1. When averaged across participants, boredom decreased, challenge increased, and frustration increased with increasing difficulty in the game. A Friedman test revealed that only the mean ratings for challenge were significantly different (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 2). A larger number of participants might yield results where each successive difficulty level is perceived as more challenging than the previous level.

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.

Table 1: Mean subjective responses for boredom, challenge, frustration, and fun for each of the difficulty levels. A response of "1" corresponded to "low" and "5" corresponded to "high". Only the ratings for challenge were significantly different.

	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 2: Wilcoxon Signed Ranks Test results for perceived challenge. Only the beginner level was perceived as significantly less 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)

4.3.2. Physiological measures

Because the subjective ratings were not consistent for each participant, 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.

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: $F_{3,12}$ =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). 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).

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 (see Figure 3). There is no simple explanation for this result, but considering that HR tends to increase with positive affect as compared to negative affect (Winton et al., 1984), 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.

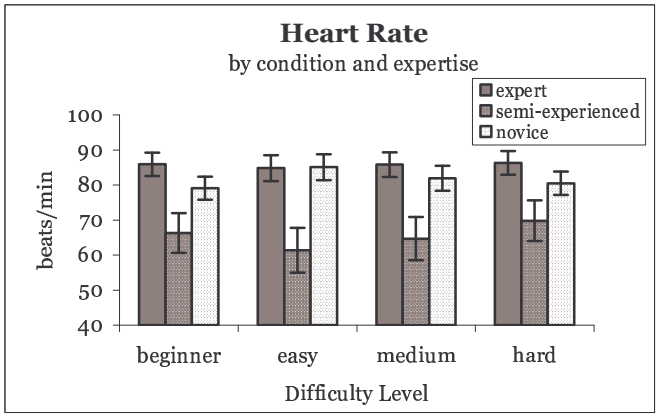

Figure 3: Heart Rate split by difficulty condition and expertise. There were no differences between conditions for experts, but there were significant differences for semi-experienced and novice players.

4.3.3. 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 3.

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 inconsistency 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 in the subsequent section.

4.4. Issues in Experiment One

There were a number of issues that impacted the results 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, thus 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.

Table 3: Significant correlations between subjective ratings and mean physiological measures for each participant. The seven 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	+	.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	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.

High resting baseline: Resting rates were sometimes higher than game play rates for some measures (e.g. HR, HRV, GSR). Anticipation and nervousness seemed to have caused the resting baselines to be artificially high. Vicente et al. (Vicente et al., 1987) 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.

Order and Interview effects: When examining the data, we noticed that the order of condition may have impacted the results. 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.

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.

As a result, considerable order effects were observed. For example, one participant's GSR signal over the course of the experiment is shown in Figure 4. 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. 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.

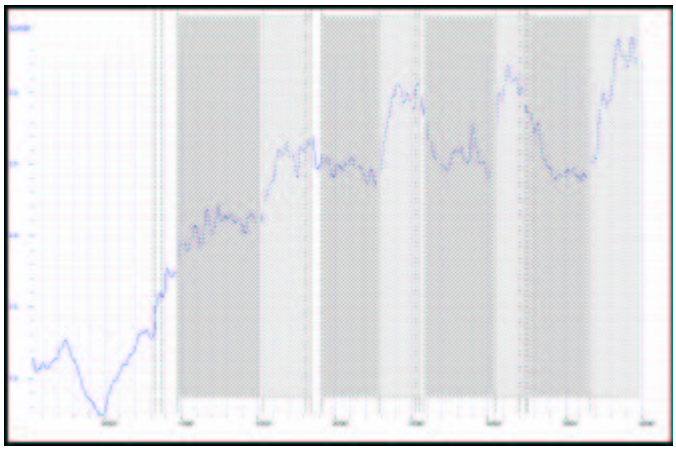

Figure 4: Participant 7's GSR signal over the course of the experiment. The areas shaded in light grey represent when the participant was being interviewed. The areas shaded in dark grey represent when the participant was playing the game.

4.5. 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.

5. Experiment Two

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 measure collaboration and engagement, 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 conditions because we have previously observed pairs (and groups) of participants playing together under a variety of collaborative conditions (Inkpen et al., 1995; Danesh et al., 2001; Mandryk et al., 2001; Scott et al., 2003). 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 main suppositions for Experiment Two are that participants will be more excited, and prefer playing against a friend than against a computer. They will have more fun, and be more engrossed in play against a friend. This preference will 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 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 our two main conjectures: that physiological measures can be used to objectively measure a player's experience with entertainment technology; and that normalized physiological measures of experience with entertainment technology will correspond to subjective reports.

5.1. Participants and setting

Ten male participants age 19 to 23 took part in the experiment. Before the experimental session, all participants filled out the same background questionnaire as in Experiment One. 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. When asked how much they liked different game genres, role-playing was the favorite, followed by strategy games.

Participants played the game in two conditions: against another player, and against the computer, and order of presentation was counterbalanced. Participants were recruited in pairs so that they would be playing against friends rather than against strangers. Participants played NHL 2003™ by EA Sports™ in both conditions (see Figure 1 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. The game settings were consistent with Experiment One. The only difference between pairs was that experienced pairs played both conditions in a higher difficulty setting than non-experienced pairs.

The experiment was conducted in a laboratory at Simon Fraser University, while all other experimental settings were consistent with Experiment One. Due to the fact that there were two players, there was no close-up view of the controller. 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 5). The video was used to determine the times of the rest periods and two play conditions, for identification of certain events (e.g. goal, fight), and for qualitative analysis, but not for quantitative video annotation. Upon arriving, participants signed a consent form. They were then fitted with the physiological sensors. One participant rested for five minutes, and then played the game against the computer. Both participants then rested for five 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 was used to help them ignore the other player in the room. They also listened to the CD when resting alone to maintain consistency. The resting period was included not for baseline comparison, but 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 5: Quadrant display including: a screen capture of the biometrics, a screen capture of the game, and the camera feed of the participants.

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. 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 five-point scale with one corresponding to "Strongly Disagree" and five corresponding to "Strongly Agree". The same technique was used to rate how challenging, easy, engaging, exciting, frustrating, and fun that particular condition was. Experiment One revealed that the physiological measurements for all participants reacted strongly to the interview process between each condition. As a result, we chose to have participants fill out questionnaires in Experiment Two using a laptop computer, followed by a five-minute rest. After completing the experiment, subjects completed a post-experiment questionnaire. We asked them to decide in retrospect which condition was more enjoyable, more fun, more exciting, and more challenging. They were also asked which condition they would choose to play in, 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.

5.2. Results and discussion

The raw physiological signals were analysed in the same manner as in Experiment One. 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.

5.2.1. 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) (χ2=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) (χ2=4.50, p=.343). The game outcome had no significant impact on any of the other subjective measures.

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 (Wilcoxon test for 2 related samples). See Table 4 for a synopsis of these results.

On the post-experiment questionnaire, when asked whether it was more enjoyable to play against the computer or a friend, all ten subjects chose playing against a friend. All ten 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. When asked if given a choice, in which condition they would choose to play, all ten 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.

Table 4: Results of condition questionnaires. Subjects were asked to rate each experience state on a five-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**

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.

5.2.2. Physiological measures

Means for the physiological data were analysed using a repeated measures multivariate analysis of variance (MANOVA) with the two play conditions as independent variables, and the five physiological signals as dependent variables. Order of presentation, and challenge group 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.

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. Overall, 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 is 36% of the signal span. Also, the partial eta-squared value of .60 reveals that 60% of the total variability in the measure can be attributed to the factor of play condition.

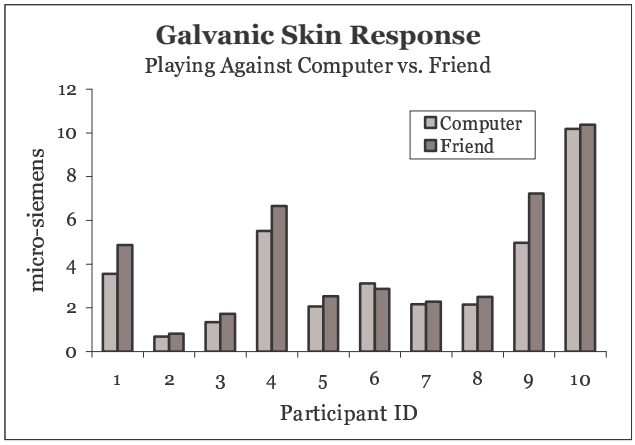

Figure 6: GSR was higher when playing against a friend as compared to playing against a computer. This pattern was seen in all players with the exception of participant 6.

In addition, when examined individually, this increase was consistent for 9 of the 10 subjects, which is a significant trend (Z=2.4, p=.017, Figure 6). 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.

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 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). The factor of condition accounts for 75% of the variability in the measure, and the increase was consistent for 9 of the 10 subjects, which is a significant trend (Z=2.7, p=.007).

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).

5.2.3. 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[1]. 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 they 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 analysed for ten seconds before scoring and fifteen seconds after scoring. There were ten instances where participants 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 7.

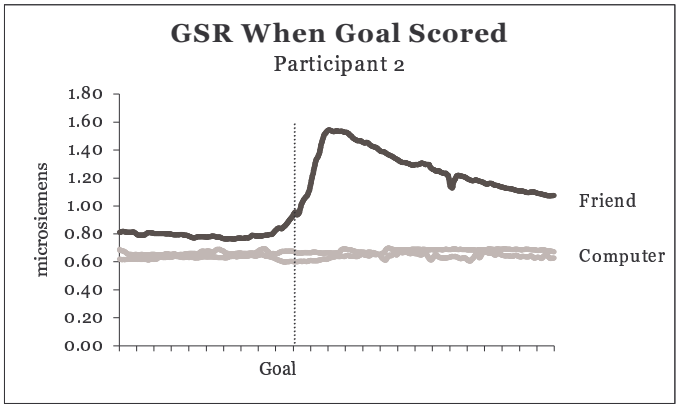

Figure 7: Participant 2's GSR response to scoring a goal against a friend and against the computer twice. Note the much larger response when scoring against a friend. Data were windowed 10 seconds prior to the goals and 15 seconds after.

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 9). 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 8.

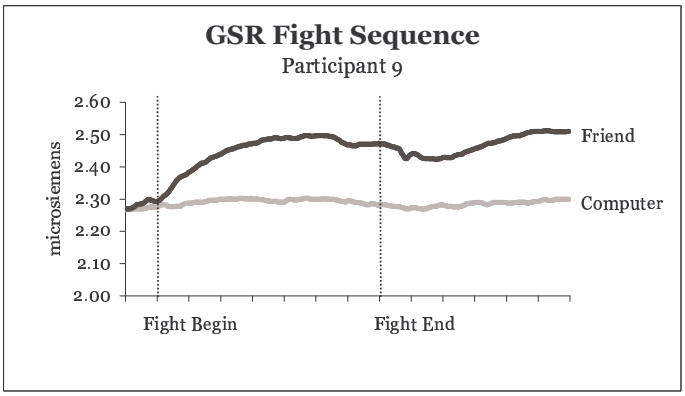

Figure 8: Participant 9's GSR response to engaging in a hockey fight playing against a friend and playing against the computer. Note that the baseline GSR for participant 9 is higher than that of participant 2 in Figure 7. This is due to individual differences, as well as differences in electrode placement and contact between participants.

Figure 9: 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.

5.2.4. Correlation of subjective responses and physiological data

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.

There was a main effect of challenge group on EMG. Those who felt that playing against the computer was more challenging had a higher mean EMG 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.

[1] 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 (Stern et al., 2001).

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, researchers have rarely correlated physiological data to other types of data. One exception is Vicente et al. (Vicente et al., 1987) who normalized HRV data by dividing the results by the resting values and subtracting this result from one.

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, 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; the difference between the conditions was divided by four, the span of a five-point scale:

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

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 inversely 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 5.

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 10). 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.

Table 5: Correlations between normalized subjective measures and normalized physiological measures. Significant correlations (2-tailed) are shaded in grey. *r* values are Pearson Correlation.

		GSR	HR	RespAmp	RespRate	EMG
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

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 ten 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, nine of the ten subjects rated the challenge of playing against a friend as the same or higher than playing against the computer in the condition questionnaires.

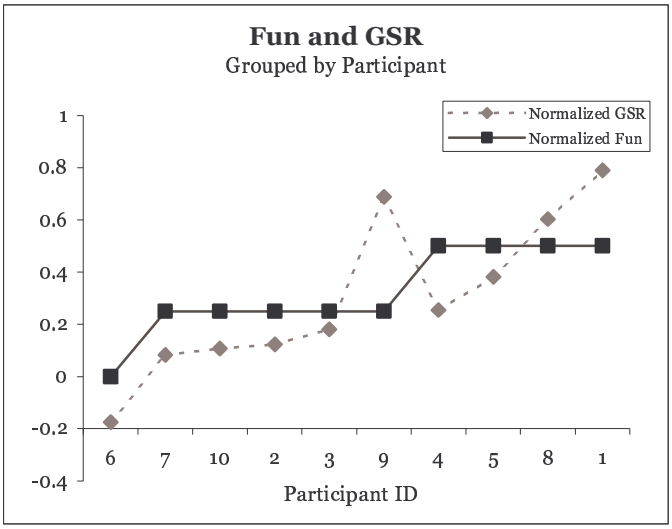

Figure 10: Normalized GSR is correlated with normalized fun (r = .70, p =.026).

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 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 (see section 5.2.1), 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.

5.3. Summary of Experiment Two

After addressing our methodological issues from Experiment One, Experiment Two tested and supported four experimental hypotheses: 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 that the differences in the participants' GSR signal in the two conditions was correlated to the differences in their subjective responses for fun and inversely with their subjective responses for frustration. We also found other correlations between the normalized subjective measures and the normalized physiological measures.

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 5.2.2 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 10 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 subjective 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.

We have been able to show that physiological responses correspond to how the players perceive the play environments. Our results reveal that GSR and EMG are higher when playing against a friend over playing against a computer. These results indicate that players are more engrossed in the game when playing against a friend, they have more fun, find it less boring, and tend to be more competitive.

6. Future Work

Once the methodological issues with collecting physiological data were addressed, Experiment Two showed that mean physiological measures show significant differences under different experimental conditions. The raised mean 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. These two experiments examined mean physiological responses, with only a few examples of examining how the physiological measures changed over time. A few examples of player's GSR reactions to goals scored and fighting in the game reveal the potential of GSR and other physiological measures to provide high-resolution, continuous measures, grounded in the context of the experience to discriminate between experiences with greater resolution than averages alone. In this paper, we graphically represented continuous responses to different game events, and looked at the magnitude of the response using the span of the physiological measure. Our future work proposes to use similar experimental manipulations to generate continuous data, from which interesting features can be extracted and utilized.

Using the collected signals, we will look for different patterns and features in the data. These features could be spikes, local minimums, local maximums, increases, decreases, the rate of change of the signal, or number of peaks per unit time. We will decide which features to investigate based on our theoretical knowledge of the physiological signals, and on inspection of empirical results. We expect that by using well-understood experimental manipulations, we will find very interesting features in the continuous, contextual physiological data. Previous annotations of the data sources have shown that distinct game events produce visible physiological responses. Our mathematical approach to the analysis (as opposed to a qualitative, annotative approach) should provide replicable results that can be extrapolated to other play situations.

Once feature-based techniques are developed, the need for a comparison system will diminish. One of our motivators for this research is that novel entertainment environments do not generally have baseline systems with which to make a comparison. If we understand how data features are indicative of player experience, we will not need control conditions. A designer or developer can evaluate their stand-alone systems.

We would also like to consider EMG on the face. Corrugator supercilii (the frowning muscle) activity and zygomaticus major (the smiling muscle) activity will be collected. Previous research has shown that smiling activity increases with positive emotions, while frowning activity (and brow furrowing in concentration) increases with negative emotions (Cacioppo et al., 2000; Partala & Surakka, 2004). As such, the EMG signals may be able to determine the valence of arousal. In fact, the examination of these three signals alone might be enough to create an approximation of an arousal-valence space (Russell et al., 1989; Lang et al., 1993), used by many researchers to classify emotion. Finally, we have recently been examining other play conditions, such as playing against a stranger, or playing against a friend over a network. This paper presented the correlation between physiological data and subjective ratings provided by the participant. Future work will address the correlation between sources of objective data (e.g. game events) and the physiological data, completing the triangulation of data sources.

7. Conclusions

The evaluation of user experience with entertainment technology is ripe for advancement. Subjective data yield valuable quantitative and qualitative results. However, when used alone, they do not provide sufficient information. Physiological measures have previously been used to evaluate productivity systems, especially to reflect a user's stress or mental effort. The application of physiological measurement and analysis to collaborative leisure technology has exciting potential.

The methodological problems that we initially experienced in collecting and analyzing physiological data revealed a number of caveats for conducting this type of research. For example, great care must be taken to avoid stimuli that affect emotional responses, other than the stimuli being investigated. Although we took many precautions in Experiment One, such as the caffeine intake, sex, and age of the participants, there were still effects that we did not predict, such as the responses generated from the interview process. In addition, we know that conducting controlled experiments with multiple users is extremely challenging due to large individual variations in physiological responses. Comparing one subject to another is not an acceptable approach, however, having each participant act as their own control, (as we did in Experiment Two), is an effective and valuable method. For scenarios where performing group analysis is desired, undertaking a normalization process is a viable option. We presented one approach to normalizing physiological data for two experimental conditions. For more than two conditions, using the resting baseline as a normalization tool would be a feasible practice. As we continue to conduct research based on physiological data, we plan to formalize these methodological approaches into a number of guidelines that researchers interested in collecting physiological data for analysis of interactive computer systems can follow to ensure legitimate data and valid results.

After addressing our methodological issues, Experiment Two tested and supported four experimental hypotheses. The confirmation of these 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.

Although these results are an encouraging progression towards user experience analysis, they have the same disadvantage as subjective results. 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.

Although we do not currently understand how the body physically responds to enhanced interaction, or increased enjoyment, a continuation of benchmark studies like this one will ultimately provide researchers with a methodology for objectively evaluating user experience with entertainment technologies. We foresee that objective evaluation, combined with current subjective techniques will provide researchers with techniques as rigorous and valuable as current methods of evaluating user performance with productivity systems.

8. References

BJÖRK, S., FALK, J., HANSSON, R.& LJUNGSTRAND, P. (2001). Pirates! Using the Physical World as a Game Board. In *Proceedings of Interact 2001*. Tokyo, Japan.

BJÖRK, S., HOLOPAINEN, J., LJUNGSTRAND, P.& MANDRYK, R.L. (2002). Introduction to Special Issue on Ubiquitous Games. *Personal and Ubiquitous Computing*, *6*: p. 358–361.

BOUCSEIN, W., (1992). *Electrodermal Activity*. The Plenum Series in Behavioral Psychophysiology and Medicine, ed. W.J. Ray New York: Plenum Press.

CACIOPPO, J.T., BERNTSON, G.G., LARSEN, J.T., POEHLMANN, K.M.& ITO, T.A., (2000). The Psychophysiology of Emotion, in *Handbook of Emotions*, J.M. Haviland-Jones, Editor. The Guilford Press: New York.

CACIOPPO, J.T.& TASSINARY, L.G. (1990). Inferring Psychological Significance From Physiological Signals. *American Psychologist*, *45*(1): p. 16-28.

DANESH, A., INKPEN, K.M., LAU, F., SHU, K.& BOOTH, K.S. (2001). Geney: Designing a collaborative activity for the Palm handheld computer. In *Proceedings of Conference on Human Factors in Computing Systems*. Seattle, WA, USA: ACM Press. p. 388-395.

EKMAN, P., (1999). Basic Emotions, in *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power, Editors. John Wiley & Sons, Ltd.: Sussex.

EKMAN, P., LEVENSON, R.W.& FRIESEN, W.V. (1983). Autonomic Nervous System Activity Distinguishes among Emotions. *Science*, *221*(4616): p. 1208-1210.

FISHER, C.& SANDERSON, P. (1996). Exploratory Data Analysis: Exploring Continuous Observational Data. *Interactions, 3* (2).

HOLMQUIST, L.E., FALK, J.& WIGSTRÖM, J. (1999). Supporting Group Collaboration with Inter-Personal Awareness Devices. *Journal of Personal Technologies*, *3*(1-2).

INKPEN, K., BOOTH, K.S., KLAWE, M.& UPITIS, R. (1995). Playing Together Beats Playing Apart, Especially for Girls.

In *Proceedings of Computer Supported Collaborative Learning*.

LANG, P.J. (1995). The Emotion Probe. *American Psychologist*, *50*(5): p. 372-385.

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

LEVENSON, R.W. (1992). Autonomic Nervous System Differences Among Emotions. *American Psychological Society*, *3*(1): p. 23-27.

MAGERKURTH, C., STENZEL, R.& PRANTE, T. (2003). STARS - A Ubiquitous Computing Platform for Computer Augmented Tabletop Games. In *Proceedings of Video Track of Ubiquitous Computing (UBICOMP'03)*. Seattle, Washington, USA.

MANDRYK, R.L., INKPEN, K.M., BILEZIKJIAN, M., KLEMMER, S.R.& LANDAY, J.A. (2001). Supporting Children's Collaboration Across Handheld Computers. In *Conference Supplement to Conference on Human Factors in Computing Systems*. p. 255-256.

MANDRYK, R.L., MARANAN, D.S.& INKPEN, K.M. (2002). False Prophets: Exploring Hybrid Board/Video Games. In *Conference Supplement to Conference on Human Factors in Computing Systems*. p. 640-641.

MARSHALL, C.& ROSSMAN, G.B., (1999). *Designing Qualitative Research. (3rd ed.)* Thousand Oaks: Sage.

NIELSEN, J., (1992). Evaluating the Thinking-Aloud Technique for Use by Computer Scientists, in *Advances in Human-Computer Interaction*, H.R. Hartson and D. Hix, Editors. Ablex Publishing Corporation: Norwood. p. 69-82.

NORMAN, D.A. (2002). Emotion and Design: Attractive things work better. *Interactions, 9* (4).

PAPILLO, J.F.& SHAPIRO, D., (1990). The Cardiovascular System, in *Principles of Psychophysiology: Physical, Social, and Inferential Elements*, L.G. Tassinary, Editor. Cambridge University Press: Cambridge. p. 456-512.

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

PICARD, R.W., (1997). *Affective Computing*. Cambridge, MA: MIT Press.

RANI, P., SIMS, J., BRACKIN, R.& SARKAR, N. (2002). Online Stress Detection using Psychophysiological Signal for Implicit Human-Robot Cooperation. *Robotica*, *20*(6): p. 673-686.

ROWE, D.W., SIBERT, J.& IRWIN, D. (1998). Heart Rate Variability: Indicator of User State as an Aid to Human-Computer Interaction. In *Proceedings of Conference on Human Factors in Computing Systems*. p. 480-487.

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): p. 493-502.

SCHEIRER, J., FERNANDEZ, R., KLEIN, J.& PICARD, R. (2002). Frustrating the User on Purpose: A Step Toward Building an Affective Computer. *Interacting with Computers*, *14*(2): p. 93-118.

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): p. 220-228.

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

SYKES, J.& BROWN, S. (2003). Affective Gaming: Measuring Emotion Through the GamePad. In *Conference Supplement to Conference on Human Factors in Computing Systems*: ACM Press. p. 732-733.

VICENTE, K.J., THORNTON, D.C.& MORAY, N. (1987). Spectral Analysis of Sinus Arrhythmia: A Measure of Mental Effort. *Human Factors*, *29*(2): p. 171-182.

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

WARD, R.D., MARSDEN, P.H., CAHILL, B.& JOHNSON, C. (2002). Physiological Responses to Well-Designed and Poorly Designed Interfaces. In *Proceedings of CHI 2002 Workshop on Physiological Computing*. Minneapolis, MN, USA.

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

WILSON, G.M. (2001). Psychophysiological Indicators of the Impact of Media Quality on Users. In *Proceedings of CHI 2001 Doctoral Consortium*. Seattle, WA, USA.: ACM Press. p. 95-96.

WILSON, G.M.& SASSE, M.A. (2000a). Do Users Always Know What's Good For Them? Utilizing Physiological Responses to Assess Media Quality. In *Proceedings of HCI 2000: People and Computers XIV - Usability or Else!* Sunderland, UK.: Springer. p. 327-339.

WILSON, G.M.& SASSE, M.A. (2000b). Investigating the Impact of Audio Degradations on Users: Subjective vs. Objective Assessment Methods. In *Proceedings of OZCHI 2000: Interfacing Reality in the New Millennium*. Sydney, Australia. p. 135-142.

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