Physiological Indicators for the Evaluation of Co-located Collaborative Play

Regan L. Mandryk
Simon Fraser University
School of Computing Science
Burnaby, BC, Canada
1-604-291-3610
rlmandry@cs.sfu.ca

Kori M. Inkpen
Dalhousie University
Faculty of Computer Science
Halifax, NS, Canada
1-902-494-1831
inkpen@cs.dal.ca

ABSTRACT

Emerging technologies offer new ways of using entertainment technology to foster interactions between players and connect people. Evaluating collaborative entertainment technology is challenging because success is not defined in terms of productivity and performance, but in terms of enjoyment and interaction. Current subjective methods are not sufficiently robust in this context. This paper describes an experiment designed to test the efficacy of physiological measures as evaluators of collaborative 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. We provide an initial step towards using physiological responses to objectively evaluate a user's experience with collaborative entertainment technology.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces- *Evaluation/methodology*

H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces- *Evaluation/Methodology, Collaborative Computing*

General Terms

Experimentation, Human Factors

Keywords

GSR, heart rate, EMG, fun, collaboration, physiology

1. INTRODUCTION

Emerging technologies in ubiquitous computing and ambient intelligence offer exciting new interface opportunities for co-located entertainment technology, as evidenced in recent growth in the number of conference workshops and research articles devoted to this topic [1, 2, 14, 16]. 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 the interaction and communication between players.

For example, we recently created a hybrid board-video game system to enhance player interaction [16]. 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. 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 [6, 15]. 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 [15].

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 [1, 10, 14]. However, evaluating the success of these new interaction techniques and environments is an open research challenge.

1.1 Evaluation of Entertainment and Collaborative Technologies

Traditionally, human-computer interaction research (HCI) has been rooted in the cognitive sciences of psychology and human factors, and in the sciences of engineering, and computer science [19]. Although the study of human cognition has made significant progress in the last decade, the notions of affect and emotion are equally important to design [19], especially when the primary goals are to challenge and entertain the user. This approach presents a shift in focus from usability analysis to human experience analysis. Traditional objective measures used for

CSCW'04, November 6-10, 2004, Chicago, IL, USA.

productivity environments, such as time and accuracy, are not 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 interested in whether our environment fosters interaction and communication between the players, creates an engaging experience, and is fun. Successful interaction techniques 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, such as speed or accuracy, the measures of success for collaborative entertainment technologies are more elusive. We want to increase interaction, enhance engagement, and create a fun experience. The current research problem lies in what metrics to use to measure engagement, interaction, and fun.

We have previously used both subjective reports and video coding as methods of evaluating our new technologies, although there is no control environment with which to make comparisons [15, 16, 27]. Subjective reporting through questionnaires and interviews is generalizable, convenient, amenable to rapid statistics and easy to administer. Some drawbacks are that questionaires are not conducive to finding complex patterns, can invade privacy, and subject responses may not correspond to the actual experience [17]. Knowing that their answers are being recorded, participants will sometimes answer what they think you want to hear, without even realizing it. Subjective ratings are cognitively mediated, and may not accurately reflect what is occurring [36]. Although many studies have shown this to be true, these results may not extend to the domain of entertainment and games, where personal preference is essential to the enjoyment of the 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 [18], which are popular for use in productivity systems cannot be effectively used with entertainment technology because of the disturbance to the player, and the impact they have on the condition itself. Using the technique retrospectively would only qualify the experience, rather than providing concrete quantitative data. In addition, the information provided by the retrospective think-aloud protocol would be reflective, not grounded in the context of the experience itself.

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 [9]. 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 [17]. Researchers must be careful to acknowledge their biases, address inter-rater reliability, and not read inferences where none are present [17]. 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 [9]. Consequently, many researchers rely on subjective data for user preference, rather than objective observational analysis.

Researchers in Human Factors have used physiological measures as indicators of mental effort and stress [30, 34]. Psychologists use physiological measures as unique identifiers of human emotions such as anger, grief, and sadness [8]. However, physiological data have not been employed to identify human experience states of enjoyment and fun. Physiological data is a high-resolution time series, responsive to player experience. Using methods like the ones presented in this paper could provide researchers with a continuous objective data source that can be used to evaluate the player experience.

Our research aims to uncover whether there are links and correlations between player's physiological states, events occurring during the collaborative experience, and subjective reported experience. These correlations would enable novel collaborative entertainment technologies to be tested and evaluated in terms of enhancing interaction and increasing engagement and fun. Based on previous research on the use of psychophysiological techniques (see Section 5), we believe that directly measuring and capturing autonomic nervous system (ANS) activity will provide researchers and developers of technological systems with direct 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 the research is 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 one experiment that we designed to test support for the two main conjectures. We recorded users' physiological, verbal and facial reactions to game technology, and applied 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 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 describe the experiment design, setting, and protocol. A presentation of the data analyses, results, and discussion follow. We familiarize the reader with related literature on physiology as a metric for evaluation in other domains. Finally, we present a look forward into the potential of using body responses as an evaluation of collaborative entertainment technologies.

2. PHYSIOLOGY AND EMOTION

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) and interbeat interval (IBI) were computed from the EKG signal, while respiration amplitude (RespAmp) and respiration rate (RespRate) were computed from the raw respiration. 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 that are used to measure GSR called the eccrine sweat glands. 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 [28]. 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 [28].

Galvanic skin response is a linear correlate to arousal [12] and reflects both emotional responses as well as cognitive activity [3]. GSR has been used extensively as an indicator of experience in both non-technical domains (see [3] for a comprehensive review), and technical domains [31, 32, 34, 35].

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 the movement utilized 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 heart rate (HR), heart rate variability (HRV), blood pressure (BP), and blood volume pulse (BVP). EKG (Electrocardiography) measures electrical activity of the heart. HR, interbeat interval (IBI), 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 [20]. HRV refers to the oscillation of the interval between consecutive heartbeats. It 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 dispatch [33] and air traffic control [24], 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 mental effort [24], but as the mental effort needed for a task increases beyond the capacity of working memory, HRV will increase.

Although there is a standard medical configuration for placement of electrodes, two electrodes placed fairly far apart will produce an EKG signal [28]. We placed 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 [28]. 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 [28]. Because respiration is closely linked to cardiac function, a deep breath can affect other measures.

Respiration is most accurately measured by gas exchange in the lungs, but the sensor technology inhibits talking and moving [28]. Instead, chest cavity expansion can be used to capture breathing activity using either a hall effect sensor, strain gauge, or a stretch sensor [28]. 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 [28]. Two electrodes are placed along the muscle of interest and a third ground is placed off axis. In isometric conditions (no movement) EMG is closely correlated with muscle tension [28], 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 [4]. 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) muscle regions are higher when emotions are mildly positive, as opposed to mildly negative [4]. These effects are stronger when averaged over a group rather than for individual analysis. In addition to emotional stress and emotional valence, EMG has been used to distinguish facial expressions and gestural expressions [28].

We used surface electrodes to detect EMG on the jaw, indicative of tension. Our previous observations of jaw EMG during computer game play have shown that jaw clenching tends to increase when participants are frustrated, or concentrating very hard. 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 [5], and although this viewpoint is too simplistic, recent evidence suggests that physiological data sources can differentiate among some emotions [13]. There are varying opinions on whether emotions can be classified into discrete, specific emotions [7], or whether emotions exist along multiple axes in space [12, 25]. Both theoretical camps have seen limited success in using physiological data to identify emotional states (see [4] for an overview). In addition to the difficulties in classifying emotions, when using physiological data sources there are methodological issues that must be addressed [22], and theoretical limitations to inferring significance [5]. Discussing these issues are beyond the scope of this paper.

3. EXPERIMENT DESIGN

To better understand how body responses can be used to create an objective evaluation methodology, and to look for support for our two main conjectures, we observed pairs of participants playing a computer game. Because this methodology is a novel approach to measure collaboration and engagement, we used an experimental manipulation designed to maximize the difference in the experience for the participant. They 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 [6, 11, 15, 27]. 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 collaborative play environments, more subtle manipulations can be explored.

Our previous studies on collaborative play, as well as the literature on physiology and emotion (see Section 2) were used to generate the following experimental hypotheses.

H1: *Participants will prefer playing against a friend to playing against a computer.*

H2: *Participants will experience higher GSR values when playing against a friend than against a computer, due to greater arousal.*

H3: *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 competition.*

H4: *The differences in the participants' GSR signal in the two conditions will correlate to the differences in their subjective responses of arousal-related measures (e.g. fun and excitement).*

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

3.1 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. Before participating in the experiment, all participants filled out a background questionnaire. 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, 9 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, 2 played every day, 7 played often, and 1 played rarely. When asked how much they liked different game genres, role-playing was the favorite, followed by strategy games (see Table 1).

Table 1: Results of game genre preference from background questionnaires. Participants rated their enjoyment on a scale from 1 to 5. Higher means indicate a stronger preference for that game genre.

	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

3.2 Play Conditions

Participants played the game in two conditions. In one condition, participants played against another player, in the other condition, they played 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 would compete against the computer before playing against their partner, while the other player would compete against the computer after playing against their partner. This was to acknowledge effects due to the order of the presentation of conditions. 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.

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

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.

3.3 Experimental Setting and Protocol

The experiment was conducted in a university laboratory. 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. Physiological data were gathered using the ProComp Infiniti system and BioGraph Software from Thought Technologies™. 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 2).

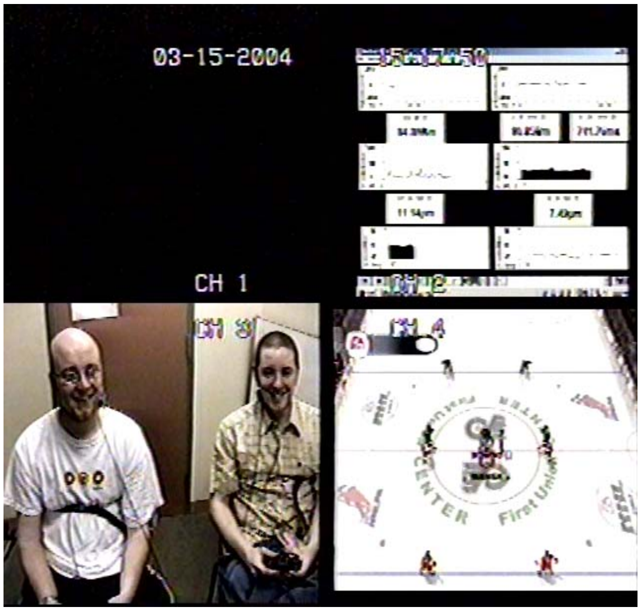

Figure 2: Quadrant display including: the screen capture of the biometrics, a screen capture of the game, and the camera feed of the participants.

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 condition 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 give us a baseline comparison, but also to allow the physiological measures to return to baseline levels prior to each condition. Our pilot experiment showed that the act of filling out the questionnaires and communicating with the experimenter can alter the physiological signals. The resting periods corrected for these effects.

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 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. 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.

3.4 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.

EKG was collected at 256 Hz, while GSR, respiration, and EMG were collected at 32 Hz. HR, IBI, RespRate, and RespAmp were 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 only one beat has 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 were also computed on the GSR data, EMG data, RespAmp data, and RespRate data.

4. 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.

4.1 Subjective Responses

H1: *Participants will prefer playing against a friend to playing against a computer.*

The chi-squared statistic was used 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) (χ^2 = 12.38, p<.02). 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) (χ^2 = 4.50, p =.343). As expected, there is some benefit when playing against a friend that is irrelevant to the game outcome. 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 < .05) to play against a computer than against a friend, but significantly more engaging (χ^2 = 4. 0, p < .05), exciting (χ^2 = 6.0, p < .02), and fun (χ^2 = 6.0, p < .02) to play against a friend than a computer (Friedman test). See Table 2 for a synopsis of these results.

Table 2: Results of condition questionnaires. Subjects were asked to rate each experience state on a scale from 1 to 5. Identifying strongly with an 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**

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 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. Those that were more challenged by their partner felt that the computer was too predictable. When asked if given a choice, 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.

4.2 Physiological Responses

Means for the physiological data were analyzed using a one-way analysis of variance to determine whether there were effects due to order of condition or outcome of the condition (win loss or tie). Neither of these factors influenced the physiological data. As a result, the physiological data for each play condition were compared using paired-samples t-tests.

H2: *Participants will experience higher GSR values when playing against a friend than against a computer, due to greater arousal.*

Our second experimental hypothesis assumed that psychological arousal would be greater when playing against a friend as compared to playing the computer. As a result, we expected that GSR would be greater when playing a friend. Overall, mean GSR was significantly higher when playing against a friend (mean = 4.19μm) as compared to playing against a computer (mean = 3.58 μm), (t_9 = 2.6, p < .03). This pattern was consistent for 9 of the 10 subjects, which is a significant trend (Z= 2.4, p < .02, see Figure 3). The one subject whose GSR did not increase 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 contributed to lower arousal and may explain his anomalous result.

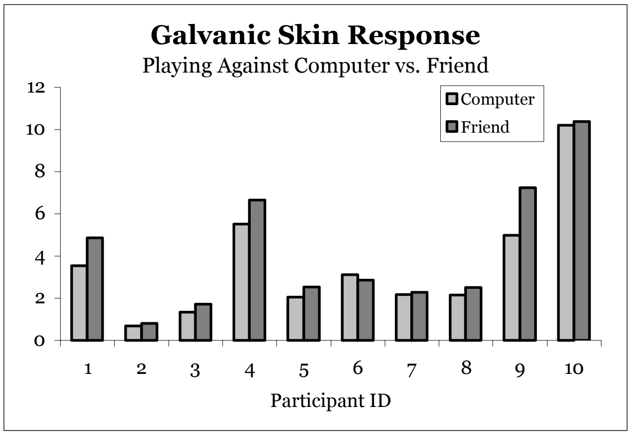

Figure 3: GSR was higher when playing against a friend as compared to playing against a computer. This pattern was seen in all players with the exception of participant 6.

H3: *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 competition.*

Our third hypothesis expected EMG activity along the jaw to be greater when playing a friend. Although we placed the surface EMG on the jaw to collect data on tension in the jaw, these results could be overshadowed by interference created from smiling and laughing. We cannot separate out these effects, to determine the EMG scores for jaw clenching alone. With this in mind, mean

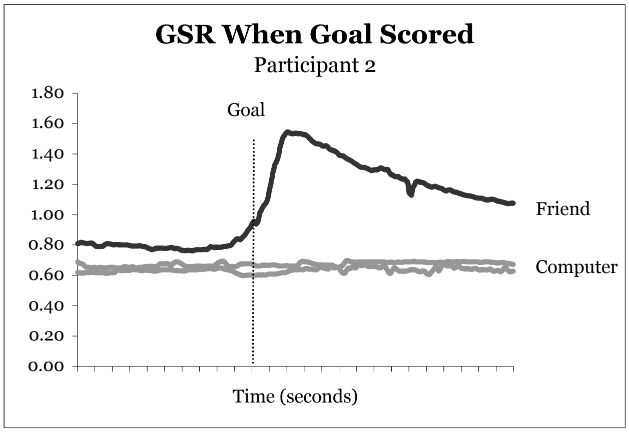

Figure 4: Participant 2's GSR response to scoring a goal against a friend and against the computer twice. Note the much larger response when scoring against a friend. Data were windowed 10 sec prior to the goals and 15 sec after.

EMG was significantly higher when playing against a friend (mean = 12.77 µV) as compared to playing against a computer (mean = 6.33 µV), (t_9 = 3.1, p < .02). This pattern was also significant for 9 of the 10 subjects, which is a significant trend (Z= 2.7, p < .01). Although these results confirm our third experimental hypothesis, they have to be interpreted with caution.

There were no significant differences in heart rate, respiratory amplitude, or respiration rate between the two play conditions. Based on our choice of experimental conditions, we didn't expect any differences in these measures, but tested them in case there were aspects of the two different experiences that we didn't acknowledge.

4.2.1 Physiological Measures as a Continuous Data Source

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, and when collected at 32 Hz, it provides a fast-response time-series metric, 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 analyzed for 10 sec before scoring and 15 sec after scoring. There were 5 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 < .005). An example of one participant's result scoring against the computer twice and against a friend once is shown in Figure 4.

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 6). Fight sequences were analyzed 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 < .03). An example of one player's response to a fight sequence against the computer and against a friend is shown in Figure 5.

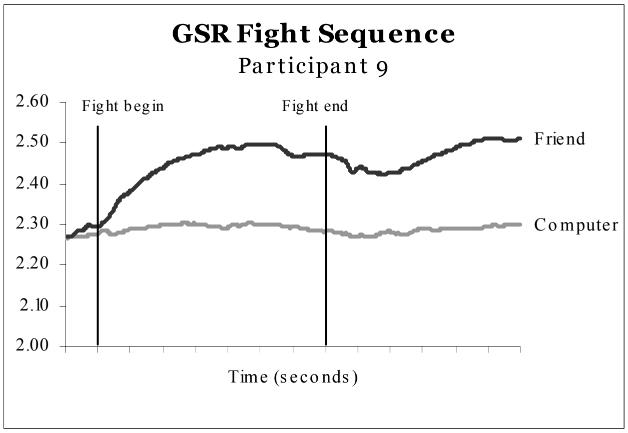

Figure 5: Participant 9's GSR response to engaging in a hockey fight with the other team while playing against a friend versus playing against the computer.

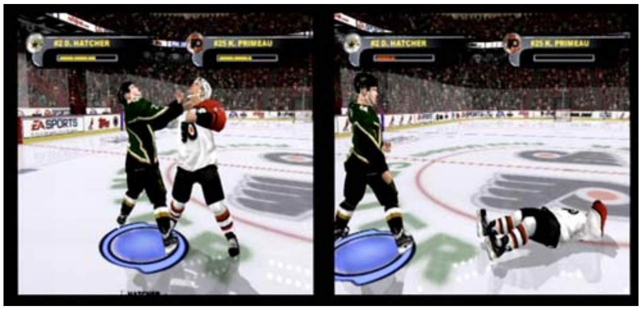

Figure 6: Fight sequence in NHL 2003 by EA Sports™. The first frame shows the players in a fight. The second frame is after the Dallas Stars ™ player won.

As discussed in the introduction, 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. Physiological data is 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.

4.3 Correlation of Subjective Data and Physiological Responses

We could not directly compare the means of the time-series data to the subjective results. Physiological data has very large individual differences, thus individual baselines have to be taken into account. Generally, one could correlate physiological results to subjective results for each individual, then determine whether these patterns were consistent across individuals. In our case, we only have two conditions, rendering this method unusable.

In order to perform a group analysis, we transformed both the physiological and subjective results into dimensionless numbers between negative one and one. For each individual, 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 inversely related to excitement. Similarly, we didn't correlate physiological measures with other physiological measures.

H4: *The differences in the participants' GSR signal in the two conditions will correlate to the differences in their subjective responses of arousal-related measures (e.g. fun and excitement).*

Since mean GSR was higher when playing against a friend, and participants also rated this condition as more fun, and exciting, we hypothesized that there may be a correlation between GSR and fun, excitement, or boredom. By themselves, the subjective and physiological results reveal that 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. We found that normalized GSR was correlated with fun ($R^2 = .72$, $p < .01$, see Figure 7). We also found that normalized GSR was inversely correlated with frustration ($R^2 = .60$, $p < .04$). Thus, the amount by which their GSR decreased when playing against the computer is proportional to the amount by which their frustration rating increased.

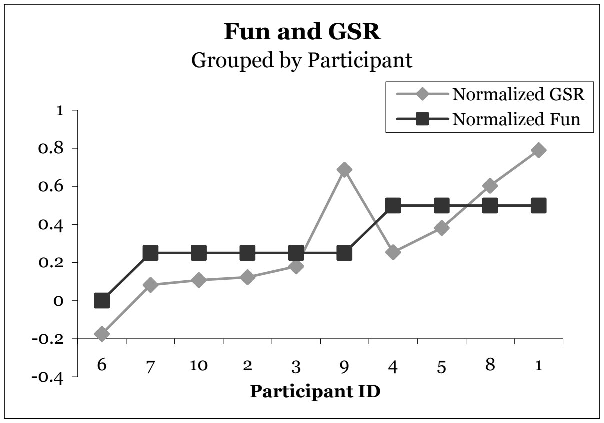

Figure 7: Normalized GSR is correlated with normalized fun ($R^2 = .72$, $p < .01$).

We also found that respiratory amplitude was correlated with challenge ($R^2 = .63$, $p < .03$). We had previously seen this same correlation when observing people playing NHL2003™ in different difficulty levels. In the current experiment, respiration amplitude increased for all ten participants when playing against a friend. Although half the participants said in the post-experiment questionnaire that playing against the computer was more challenging, 9 of the 10 subjects rated the challenge of playing against a friend as the same or higher than playing against the computer. 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, this result needs to be interpreted with caution.

5. 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 computer system has not been judged 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 [33].

Wastell and Newman [33] 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) [34-36] 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 [34]. Using 3 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 [35], 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 [36], 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. [31, 32] 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. [30] normalized heart rate variability (HRV) to a ratio between 0 and 1. 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 [23, 24, 30].

Partala and Surakka [21] and Scheirer et al. [26] 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 [29] 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.

6. FUTURE WORK

Our goal is to create a methodology for the evaluation of people's experience with collaborative entertainment technology. This paper presents initial research testing the efficacy of physiological measures as evaluators of co-located collaborative play. More steps are required to create a methodology for evaluation.

We presented results using galvanic skin response. GSR is a very responsive, salient measure that has been well studied and well documented. There are other physiological indicators that also may be relevant to collaborative play. For example, we collected EMG on the jaw to indicate tension, but our results were muddied by interference from smiling and laughing. By collecting EMG on the forehead and cheeks, we may be able to automatically detect when the participant is smiling or laughing [28]. This would provide another useful automatic measure, replacing hours of manual video analysis, or expensive facial expression recognition software. EMG of the face would also help differentiate between increases in arousal due to positive or negative activation.

Heart rate variability is another well studied measure, for measuring metal effort and stress [30]. We did not perform a heart rate variability analysis on this data, because the spectral analysis algorithm uses a five-minute time window and some of our game conditions lasted less than five minutes. Collecting longer samples in each condition and performing HRV analysis may tell us in which condition the participants exerted more mental effort.

The GSR signal revealed 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. 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 calculated the magnitude of the response. In the next phase of the research, we plan to manipulate game events and hope to take advantage of the high-resolution, contextual nature of physiological data to provide an objective, *continuous* measure of player experience.

We have compared physiological data to data from subjective reporting. Current objective methods of analysis include video coding of facial expressions, gestures, and verbal comments. We would like to compare the objective physiological results to objective data gathered through video analysis.

With a validated methodology, more subtle experimental manipulations can be explored, answering outstanding research questions in this domain. For example, what kind of entertainment experiences do ubiquitous and ambient technologies provide? And do the user's concerns with privacy overshadow the play experience? How can technologies enhance communities, both co-located and online? These questions cannot be effectively answered by subjective reporting alone. We would also like to extend the methodology to evaluate novel interaction methods and environments where suitable comparison systems do not exist.

7. CONCLUSIONS

The evaluation of enhanced interaction provided by collaboration technology, and the evaluation of fun and engagement with entertainment technology are both areas ripe for advancement. 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.

Our experiment tested and confirmed 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 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 collaborative entertainment technology. We foresee that objective evaluation, combined with current subjective techniques will provide researchers with techniques as rigorous and valuable as current methods of evaluating productivity systems.

8. ACKNOWLEDGMENTS

Thanks to Tom Calvert, Kelly Booth, and Stella Atkins for discussion. Also thanks to Thecla Schiphorst, Alan Boykiw, John Riedl, SFU Surrey, Gruvi Lab, Imager Lab, and EA Sports.

9. REFERENCES

[1] Björk, S., Falk, J., Hansson, R., and Ljungstrand, P. (2001). Pirates! Using the Physical World as a Game Board. In *Proceedings of Interact 2001*. Tokyo, Japan.

[2] Björk, S., Holopainen, J., Ljungstrand, P., and Mandryk, R.L. (2002). Introduction to Special Issue on Ubiquitous Games. *Personal and Ubiquitous Computing*, *6*: p. 358–361.

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

[4] Cacioppo, J.T., Berntson, G.G., Larsen, J.T., Poehlmann, K.M., and Ito, T.A., (2000). The Psychophysiology of Emotion, in *Handbook of Emotions*, J.M. Haviland-Jones, Editor. The Guilford Press: New York.

[5] Cacioppo, J.T. and Tassinary, L.G. (1990). Inferring Psychological Significance From Physiological Signals. *American Psychologist*, *45*(1): p. 16-28.

[6] Danesh, A., Inkpen, K.M., Lau, F., Shu, K., and Booth, K.S. (2001). Geney: Designing a collaborative activity for the Palm handheld computer. In *Proceedings of Conference on Human Factors in Computing Systems (CHI 2001)*. Seattle, WA, USA: ACM Press. p. 388-395.

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

[8] Ekman, P., Levenson, R.W., and Friesen, W.V. (1983). Autonomic Nervous System Activity Distinguishes among Emotions. *Science*, *221*(4616): p. 1208-1210.

[9] Fisher, C. and Sanderson, P. (1996). Exploratory Data Analysis: Exploring Continuous Observational Data. *Interactions, 3* (2).

[10] Holmquist, L.E., Falk, J., and Wigström, J. (1999). Supporting Group Collaboration with Inter-Personal Awareness Devices. *Journal of Personal Technologies*, *3*(1-2).

[11] Inkpen, K., Booth, K.S., Klawe, M., and Upitis, R. (1995). Playing Together Beats Playing Apart, Especially for Girls. In *Proceedings of Computer Supported Collaborative Learning (CSCL '95)*.

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

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

[14] Magerkurth, C., Stenzel, R., and 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.

[15] Mandryk, R.L., Inkpen, K.M., Bilezikjian, M., Klemmer, S.R., and Landay, J.A. (2001). Supporting Children's Collaboration Across Handheld Computers. In *Conference Supplement to Conference on Human Factors in Computing Systems (CHI 2001)*. Seattle, WA, USA. p. 255-256.

[16] Mandryk, R.L., Maranan, D.S., and Inkpen, K.M. (2002). False Prophets: Exploring Hybrid Board/Video Games. In *Conference Supplement to Conference on Human Factors in Computing Systems (CHI 2002)*. p. 640-641.

[17] Marshall, C. and Rossman, G.B., (1999). *Designing Qualitative Research. (3rd ed.)* Thousand Oaks: Sage.

[18] 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.

[19] Norman, D.A. (2002). Emotion and Design: Attractive things work better. *Interactions, 9* (4).

[20] Papillo, J.F. and 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.

[21] Partala, T. and Surakka, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with Computers*, *16*: p. 295-309.

[22] Picard, R.W., (1997). *Affective Computing.* Cambridge, MA: MIT Press.

[23] Rani, P., Sims, J., Brackin, R., and Sarkar, N. (2002). Online Stress Detection using Psychophysiological Signal for Implicit Human-Robot Cooperation. *Robotica*, *20*(6): p. 673-686.

[24] Rowe, D.W., Sibert, J., and 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 (CHI '98)*.

[25] Russell, J.A., Weiss, A., and 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.

[26] Scheirer, J., Fernandez, R., Klein, J., and Picard, R. (2002). Frustrating the User on Purpose: A Step Toward Building an Affective Computer. *Interacting with Computers*, *14*(2): p. 93-118.

[27] Scott, S.D., Mandryk, R.L., and Inkpen, K.M. (2003). Understanding Children's Collaborative Interactions in Shared Environments. *Journal of Computer Assisted Learning*, *19*(2): p. 220-228.

[28] Stern, R.M., Ray, W.J., and Quigley, K.S., (2001). *Psychophysiological Recording. (2nd ed.)* New York: Oxford University Press.

[29] Sykes, J. and Brown, S. (2003). Affective Gaming: Measuring Emotion Through the GamePad. In *Conference Supplement to Conference on Human Factors in Computing Systems (CHI 2003)*. Ft. Lauderdale, FA, USA: ACM Press. p. 732-733.

[30] Vicente, K.J., Thornton, D.C., and Moray, N. (1987). Spectral Analysis of Sinus Arrhythmia: A Measure of Mental Effort. *Human Factors*, *29*(2): p. 171-182.

[31] Ward, R.D. and Marsden, P.H. (2003). Physiological responses to different WEB page designs. *International Journal of Human-Computer Studies*, *59*(1/2): p. 199-212.

[32] Ward, R.D., Marsden, P.H., Cahill, B., and Johnson, C. (2002). Physiological Responses to Well-Designed and Poorly Designed Interfaces. In *Proceedings of CHI 2002 Workshop on Physiological Computing*. Minneapolis, MN, USA.

[33] Wastell, D.G. and Newman, M. (1996). Stress, control and computer system design: a psychophysiological field study. *Behaviour and Information Technology*, *15*(3): p. 183-192.

[34] 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.

[35] Wilson, G.M. and Sasse, M.A. (2000). 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.

[36] Wilson, G.M. and Sasse, M.A. (2000). 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.