A Continuous and Objective Evaluation of Emotional Experience with Interactive Play Environments

Regan L. Mandryk, M. Stella. Atkins
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
rlmandry@cs.sfu.ca, stella@cs.sfu.ca

Kori M. Inkpen
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada
inkpen@cs.dal.ca

ABSTRACT
Researchers are using emerging technologies to develop novel play environments, while established computer and console game markets continue to grow rapidly. Even so, evaluating the success of interactive play environments is still an open research challenge. Both subjective and objective techniques fall short due to limited evaluative bandwidth; there remains no corollary in play environments to task performance with productivity systems. This paper presents a method of modeling user emotional state, based on a user's physiology, for users interacting with play technologies. Modeled emotions are powerful because they capture usability and playability through metrics relevant to ludic experience; account for user emotion; are quantitative and objective; and are represented continuously over a session. Furthermore, our modeled emotions show the same trends as reported emotions for fun, boredom, and excitement; however, the modeled emotions revealed differences between three play conditions, while the differences between the subjective reports failed to reach significance.

Author Keywords
Emotion, play, games, fun, evaluation methodology, physiology, GSR, EMG, HR, fuzzy logic

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

INTRODUCTION
Emerging technologies in ubiquitous computing offer exciting new interface opportunities for entertainment technology, as evidenced in a recent growth in the number of conference workshops and research articles devoted to this topic (see [2, 16]). As researchers develop novel play environments, computer and console game markets continue to grow rapidly, outperforming the film industry in terms of total revenues in many regions [1]. Although technology can support compelling interactive play experiences and enhance interaction and communication between players, evaluating the success of interactive play environments is an open research challenge.

Human-computer interaction research (HCI) has been rooted in the cognitive sciences of psychology and human factors, in the applied sciences of engineering, and in computer science [22]. Although the study of human cognition has made significant progress in the last decade, the idea of emotion, which is equally important to design [22], 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 task performance, are not applicable 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, we are interested in what kind of emotional experience is provided by the play technology and environment [23]. Although traditional usability measures may still be relevant, they are subordinate to the emotional experiences resulting from interaction with the play technology and with other players in the environment.

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

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

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

This paper describes why we need such an approach; how we designed a new evaluative methodology; and how to apply this methodology for the evaluation of interactive entertainment technologies.

Evaluation of entertainment technologies

Current methods of evaluating entertainment technologies include both subjective and objective techniques. The most common methods are subjective self-reports through questionnaires, interviews, and focus groups [11] and objective reports through observational video analysis [14].

Subjective reporting through questionnaires and interviews is generalizable, convenient, and amenable to rapid statistical analysis. Some drawbacks of questionnaires and surveys are that they are not conducive to finding complex patterns, and subject responses may not correspond to the actual experience [20, 35]. Subjective techniques are good approaches to understanding the *attitudes* of the users, but subjects are bad at self-reporting their *behaviours* in game situations [23]. In addition, participants' reaction to new play environments might be skewed by the novelty of the entertainment technologies.

Using video to code gestures, body language, facial expressions and verbalizations, is a rich source of data. However, coding observational data as an indicator of human experience is a lengthy and rigorous process that needs to be undertaken with great care to avoid biasing the results [20]. The main drawback of observational video analysis is the enormous time commitment. The analysis time to data sequence time ratio (AT:ST) typically ranges from 5:1 to 100:1 [10]. There are a few consulting firms that specialize in observational analysis of entertainment technologies [14]; however, many researchers rely on subjective data for user preference, rather than objective observational analysis.

Standard discount usability methods, such as heuristic evaluation, are useful for uncovering usability issues within play environments; however, there has been minimal research on using heuristics to evaluate the playability of an entertainment technology [7, 31], or to evaluate the impact of emerging technologies. Most importantly, these discount methods do not involve actual users, but are administered by usability specialists. When research involves incorporating novel technologies into a play experience, there are no "experts". At this point, experts can only guess how the technologies will impact users.

Think-aloud techniques [21] cannot effectively be used with entertainment technology because of the disturbance to the player, and the impact they have on game play. To avoid disrupting the player during the game, researchers can employ a *retrospective* think-aloud technique. Although informative, this technique qualifies the experience, rather than providing concrete quantitative data. In addition, retrospective think-aloud does not occur within the context of the task, but in reflection of the task.

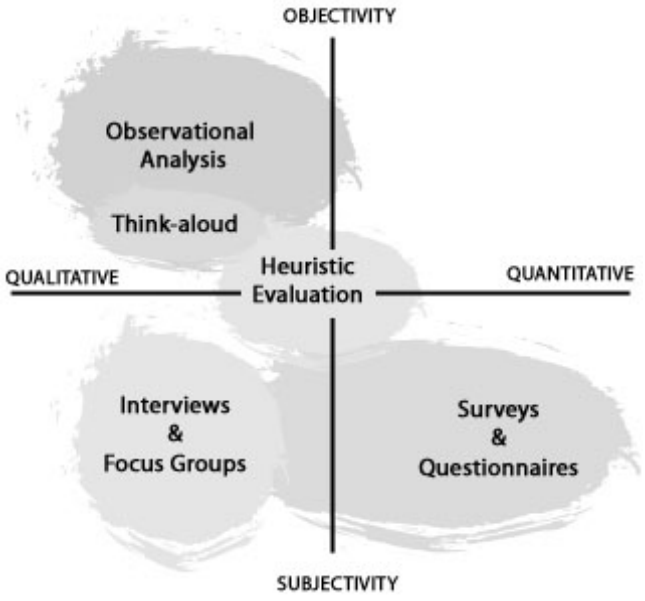

Figure 1: Current methods for evaluating entertainment technologies. Evaluators have a lot of choice, but there is a knowledge gap in the quantitative-objective quadrant. Heuristic evaluation can be quantitative since experts can provide ratings for how well software adheres to heuristics. Although observational analysis can be used for quantitative or qualitative results, it is not used quantitatively to evaluate play due to the time commitment and required expertise.

Traditional evaluation methods have been adopted, with some success, for quantitative-subjective, qualitative-subjective, and qualitative-objective assessment of play technologies. Metrics of task performance are used for quantitative-objective analysis of productivity systems, but task performance is not relevant to play [23]. As such, there is a knowledge gap for quantitative-objective evaluation of play technologies (see Figure 1). In addition, the described techniques all suffer from low evaluative bandwidth (the number of data points provided per unit time). Subjective techniques only generate data when a question is asked, and interrupting game play to ask a question is too disruptive. Heuristics also give an overview, rather than examining change over time. Using observational analysis, researchers can identify numerous events within a play session; however, the analysis is generally event-based (e.g. participant is smiling now), rather than continuous (e.g. percentage of full smile for every point in time).

Researchers in human factors have used physiological measures as indicators of mental effort and stress [32]. Psychologists use physiological measures to differentiate human emotions such as anger, grief, and sadness [9]. However, physiological data have not been employed to identify a user's emotional states such as fun and excitement when engaged with entertainment technologies. Based on previous research on the use of psychophysiological techniques, we believe that capturing, measuring, and analyzing autonomic nervous system (ANS) activity will provide researchers and developers of

technological systems with access to the emotional experience of the user. Used in concert with other subjective and/or qualitative evaluation methods, researchers can triangulate data sources and form a complex, detailed account of user experience.

We designed an experiment to create and evaluate a model of user emotional state when interacting with play technologies. We record users' physiological, verbal and facial reactions to game technology, and apply post-processing techniques to objectively and continuously measure emotional state, hence filling the knowledge gap in the objective-quantitative quadrant of Figure 1. Our ultimate goal is to create a methodology for the objective evaluation of entertainment technology, as rigorous as current methods for productivity systems, providing more choice and robustness for evaluators.

PHYSIOLOGICAL METRICS FOR EVALUATION

Researchers in the domain of human factors have 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. There are many examples of the use of physiological metrics in the domain of human factors (see [19] for an overview).

To provide an introduction for readers unfamiliar with physiological measures, we briefly introduce the measures used, describe how these measures are collected, and explain their inferred meaning. Based on previous literature, we chose to collect galvanic skin response (GSR), electrocardiography (EKG), and electromyography of the face ($EMG_{smiling}$ and $EMG_{frowning}$). Heart rate (HR) was computed from the EKG signal. The measures we used will each be described briefly including reference to how they have previously been used in technical domains.

Galvanic skin response

GSR is a measure of the conductivity of the skin. There are specific sweat glands (eccrine glands) that cause skin 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 [30]. For example, many people have cold clammy hands when they are nervous. In fact, subjects do not have to even be sweating on the palms of the hands or soles of the feet to see differences in GSR 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 [30].

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 [33-35]. We measured GSR using surface electrodes sewn in Velcro straps placed around two fingers on the same hand.

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. Electrocardiograms (EKG) measure electrical activity of the heart, and HR, IBI, and HRV can be computed from EKG.

HR reflects emotional activity. It has been used to differentiate between positive and negative emotions with further differentiation using finger temperature [24, 36]. 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 [27].

To collect EKG, we placed three pre-gelled surface electrodes in the standard configuration of two electrodes on the chest and one electrode on the abdomen.

Electromyography

Electromyography (EMG) measures muscle activity by detecting surface voltages that occur when a muscle is contracted [30]. In isometric conditions (no movement) EMG is closely correlated with muscle tension [30]. 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 (*corrugator supercilii*: frown muscle) region is lower and EMG activity over the cheek (*zygomaticus major*: smile muscle) is higher when emotions are mildly positive, as opposed to mildly negative [4].

We used surface electrodes to detect smiling activity ($EMG_{smiling}$) from *zygomaticus major* activation and frowning activity ($EMG_{frowning}$) from *corrugator supercilii* activation. The disadvantage of using surface electrodes is that the signals can be muddied by other facial muscle activity, such as talking. Needles are an alternative to surface electrodes that minimize interference, but were not appropriate for our experimental setting.

Use of physiological metrics in HCI

Physiological metrics have only recently been used in the domain of HCI. Researchers have used GSR and cardiovascular measures to examine subject response to video and audio degradations in video conferencing software [34, 35], and to investigate user response to well- and ill- designed web pages [33]. HRV has been used as an indicator of mental effort and stress when interacting with simulators [27, 32] and to distinguish between attentive states of a user [6]. Partala and Surakka [25] and Scheirer et al. [29] both used pre-programmed mouse delays to

intentionally frustrate a computer user. Partala and Surakka measured EMG activity on the face in response to affective audio intervention, while Scheirer et al. applied Hidden Markov Models to detect states of frustration.

Our previous work has examined physiological responses to different interactive play environments [18, 19]. We showed that GSR and EMG of the jaw were higher when playing against a friend, over playing against a computer, and we found many correlations between normalized physiological activity and normalized subjective measures, including strong correlations between GSR and fun, and EMG and challenge. We also showed how physiological measures provide a rich, continuous, and objective source of information about user experience with interactive entertainment technologies. Based on these results, we believe that physiological metrics can be used to model user emotional experience when playing a game; providing continuous and objective metrics of emotion.

IDENTIFYING EMOTIONS

There has been a long history of researchers attempting to use physiological data to identify emotional states. William James first speculated that patterns of physiological response could be used to recognize emotion [4], and although this viewpoint is too simplistic, recent evidence suggests that physiological data sources can differentiate among some emotions [9, 15]. Opinions vary on whether emotions can be classified into discrete emotions [8], or whether emotions exist along multiple axes [12, 28]. Both perspectives have seen limited success in using physiology to identify emotional states [4]. The arousal-valence space (AV space) used by Lang [12] classifies emotions in 2-D space defined by arousal and valence (pleasure). Using pictures as stimuli, Lang and colleagues mapped individual pictures to emotions as defined by the space.

Russell et al. [28] also used an arousal-valence space to create the Affect Grid. Based on their circumplex model of emotion, the Affect Grid is a tool to quickly assess affect along dimensions in AV space. Subjects place checkmarks in the squares of the grid, as a response to different stimuli (see Figure 2). One problem with the AV space method of classifying mood is that arousal and valence may not be independent and can impact each other. For example, Lang et al. [13] had difficulty finding images that represent the extreme regions of the unpleasant/calm quadrant. It seems that if an image is truly unpleasant, it cannot also be calm, suggesting some interplay between these two axes.

In addition to the difficulties in classifying emotions, when using physiological data sources there are methodological issues that must be addressed [26], and theoretical limitations to inferring significance [5]. Discussing these issues are beyond the scope of this paper.

USER STUDY

We conducted a study to inform the design of a continuous model of emotion, based on physiological responses. The

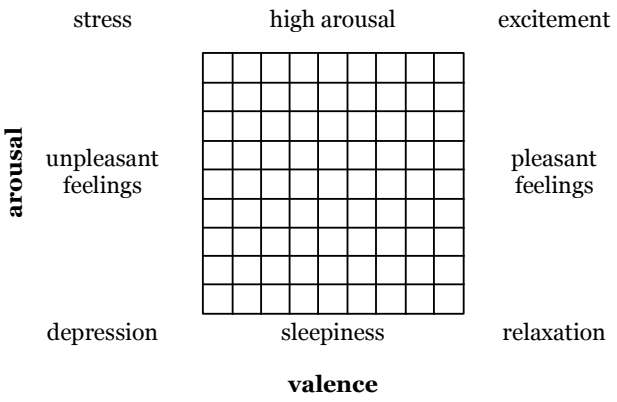

Figure 2: The Affect Grid: Based on the circumplex model of emotion, the affect grid allows for a quick assessment of mood as a response to stimuli in arousal-valence space [28].

participants played a game in three conditions: against a co-located friend, against a co-located stranger, and against the computer. As with our previous work, we were not interested in whether there was a difference between playing against a friend, a stranger, or a computer. We have observed many groups of people playing with interactive technologies, and we know that these three play conditions yield very different play experiences; rather, we were interested in whether our model of emotion could detect the differences between the conditions.

Participants

Twenty-four male participants age 18 to 27 took part in the experiment. Before the experiment, all participants filled out a background questionnaire, 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.

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

All participants were frequent computer users. When asked to rate how often they used computers, all 12 subjects used them every day. Participants were also frequent gamers, playing either computer games or console games regularly.

Play conditions

Participants played the game in three conditions: against a co-located friend, against a co-located stranger, and against the computer. Order of the presentation of the conditions was fully counterbalanced. Participants played NHL 2003 by EA Sports in all conditions (see Figure 3). Six of the pairs were very experienced or somewhat experienced with the game, three pairs were neutral in their experience, while

the other three pairs were somewhat inexperienced. The stranger remained constant for all participants, and was a 29 year-old male gamer, who was instructed to match each participant's level of play to the best of his ability.

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

Experimental setting and protocol

The experiment was conducted in a laboratory at Simon Fraser University. NHL 2003 was played on a Sony PS2, and viewed on a 36" television. A camera captured both of the players, their facial expressions and their use of the 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 3). The recording also contained audio of the participants' comments from a boundary microphone, and audio output from the game.

Physiological data were gathered using the ProComp Infiniti system and sensors, and BioGraph Software from Thought Technologies. Based on previous literature, we chose to collect galvanic skin response (GSR), electrocardiography (EKG), and electromyography of the face ($EMG_{smiling}$ and $EMG_{frowning}$). Heart rate (HR) was computed from the EKG signal. We only collected physiological data for the participant, not for the friend or stranger. To maintain the perception that both players were participants in the experiment, we treated both players as if their physiological signals were being collected. We fitted both players with sensors, tested the sensor placement to ensure that the signals were good, and plugged the extra sensors into ports on the back of the unit.

Upon arriving, participants signed a consent form. They were then fitted with the physiological sensors. Before each experimental condition, participants rested for 5 minutes while listening to a CD containing nature sounds. The resting period allowed the physiological measures to return to baseline levels prior to each condition. In prior experiments we saw that the act of filling out the questionnaires and communicating with the experimenter altered the physiological signals [19]. The resting periods corrected for these effects.

After each condition, subjects rated the condition using a Likert Scale. They were asked to consider the statement,

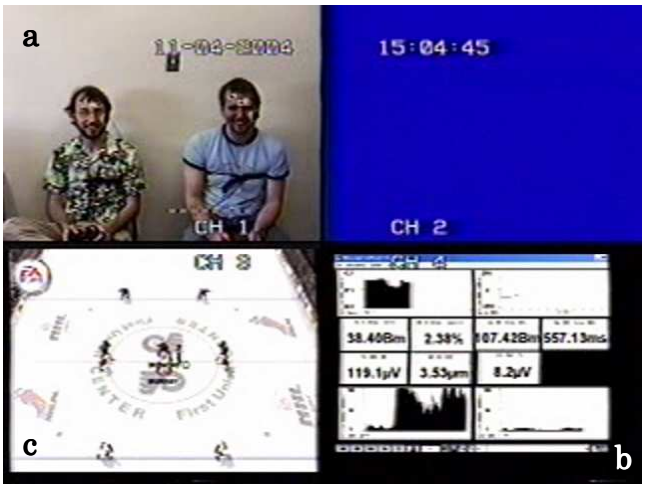

Figure 3: Quadrant display: a) camera feed of the participants, b) screen capture of the biometrics, c) screen capture of the game, audio of the game, and audio of the participants' comments.

"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, exciting, frustrating, and fun the condition was. The html-based questionnaire was filled out using a laptop computer to reduce the physiological impact of communicating with the experimenter [19]. After completing the experiment, subjects completed a post-experiment questionnaire. We asked them to decide in retrospect which condition was most fun, most exciting, and most challenging.

Data analyses

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

Each data signal was smoothed with a moving average window of 4 frames (0.125 seconds), with the exception of GSR, which was filtered using a 5-second window [3]. We then normalized each signal into a percentage between 0 and 100. There are very large individual differences associated with physiological data, and normalizing the data is necessary in order to perform a group analysis. We transformed each sample into the percentage of the span of that particular signal, for that particular participant across all three conditions. Using GSR as an example, a global minimum and maximum GSR were obtained for each participant using all three conditions and the rest period,

and the same global values were used for normalizing within each condition.

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

BUILDING THE MODEL OF EMOTION

We used the normalized GSR, HR, $EMG_{smiling}$, and $EMG_{frowning}$ signals as inputs to a fuzzy logic model. To generate values for user emotion, we modeled the data in two parts. First, we computed arousal and valence values from the normalized physiological signals, then used these arousal and valence values to generate emotion values for boredom, challenge, excitement, frustration, and fun. To generate a model of emotion, we used half of the participants (one for each play condition order), reserving the other six participants for validation of the model.

Details of how the fuzzy system was designed (the development, implementation, and comparison of the output of the fuzzy logic models to a manual approach) can be found in [17]. The current paper presents a high-level description of the model, the comparison of the model to reported emotion, and its potential use in HCI evaluations.

Modeling AV space

To make use of the continuous nature of physiological data, we used the complete time series for each input. As such, we were able to generate a new time series of the participant's experience in AV space, rather than having only one data point for an entire condition (e.g. mean).

Our model of physiology to AV space had four inputs (GSR, HR, $EMG_{smiling}$, and $EMG_{frowning}$) and two outputs (arousal and valence) (see Figure 4). Inputs were normalized signals (0-100), while outputs were percentages of the possible maximum (0-100) value for arousal and valence. For each input signal, the membership functions were generated using characteristics of that particular signal over all participants and conditions. The 22 rules were grounded in the theory of how the physiological signals relate to the psychological concepts of arousal and valence. GSR correlates with arousal, and increasing GSR was mapped to increasing arousal. The extreme high and low levels of GSR were modulated by HR data; if HR was contradictory, arousal was altered, otherwise arousal was maintained. Valence increased with increasing levels of $EMG_{smiling}$, and decreased with increasing levels of $EMG_{frowning}$. A full discussion of the membership functions and rules for the model can be found in [17], while Figure 5 shows the surfaces generated from the model.

Modeling emotion from AV space

To make the most of the rich, continuous physiological data, we modeled the entire AV space time series, creating continuous metrics of emotional experience.

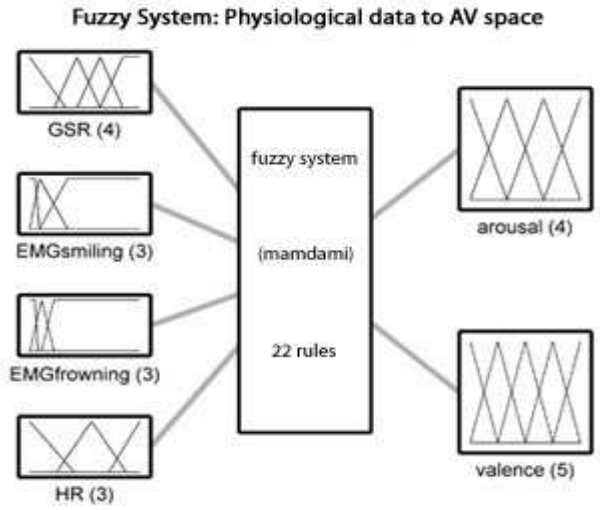

Figure 4: Modeling arousal and valence from physiological data. The number of membership functions applied to that input or output follows the input /output labels. The system used 22 rules to transform the 4 inputs into the 2 outputs.

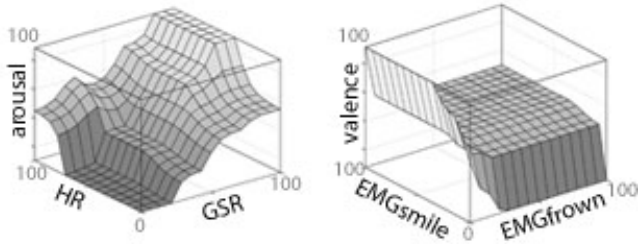

Figure 5: Surfaces depicting how GSR, HR, $EMG_{smiling}$, and $EMG_{frowning}$ are converted into arousal and valence.

Using the Affect Grid [28], developed from the circumplex model of emotion (Figure 2), we translated our arousal and valence values from the first model into a language of emotion. Five emotions were modeled: boredom, challenge, excitement, frustration, and fun. These are the same five emotions that participants rated after each play condition. As such, our AV to emotion model (see Figure 6) had two inputs (arousal and valence), and five outputs (boredom, challenge, excitement, frustration, and fun). Membership functions for the outputs, and the rules were generated by dividing emotions into four states based on AV space: very low, low, medium, and high (see Figure 7). A comprehensive discussion of the membership functions and rules for the model can be found in [17]. Inputs and outputs were represented as percentages of the possible maximum.

COMPARISON OF MODEL TO SUBJECTIVE DATA

To analyze the effectiveness of our model, we used data gathered from the six subjects not used in the generation of the model. Data were smoothed and normalized using the previously described method. Both models were applied to the data and the time series for each emotion were averaged to compare modeled emotion to the subjective responses.

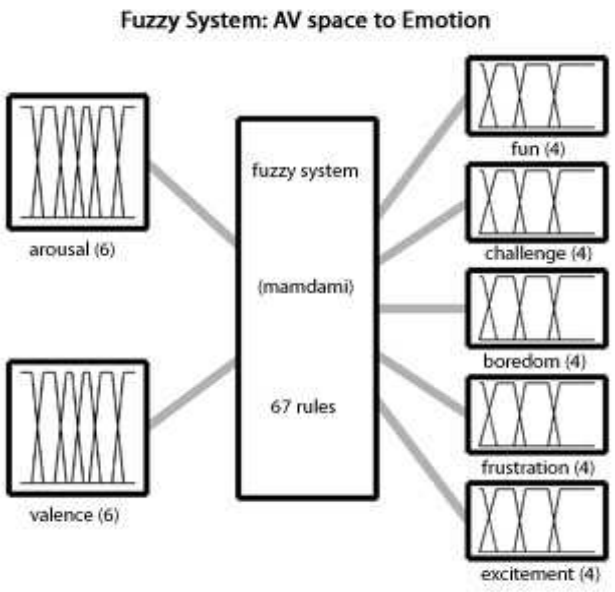

Figure 6: Modeling emotion from arousal and valence. The number of membership functions applied to that input or output follows the input /output labels. The system used 67 rules to transform the 2 inputs into the 5 outputs.

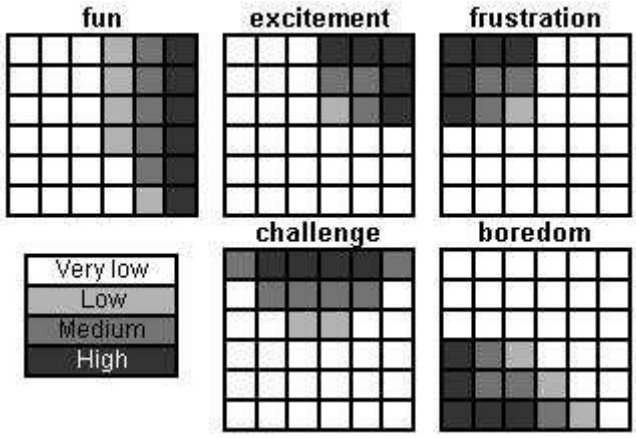

Figure 7: Our representation of levels of emotion in arousal-valence space. The x axis indicates increasing valence, while the y axis indicates increasing arousal.

Modeled emotion

Mean modeled emotions (represented as a percentage) were analyzed using a repeated measures MANOVA with the five emotions as dependent measures, and play condition as a within-subjects factor. Play condition significantly impacted fun and excitement, but not frustration, boredom, or challenge (see Table 1, Figure 8). Post-hoc analysis revealed that players were having more fun when playing against a friend than when playing against a stranger or a computer, and that playing against a stranger was more fun than playing against a computer. Playing against a friend was more exciting than playing against the computer, while playing against a stranger was marginally more exciting than playing against the computer.

Reported emotion

Participants were asked to rate the boredom, challenge, excitement, frustration, and fun of each condition on a 5-point scale. Friedman tests for 3-related samples revealed no differences between conditions (see Table 2, Figure 9).

Comparing modeled and reported emotion

Although there were no subjective differences between conditions, plotting the means reveals that there were definite trends (see Figure 9). Furthermore, plotting the modeled emotion means reveals the same trends for boredom, excitement, and fun (see Figure 8).

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

There was a correlation for challenge (rho=.99, p<.001), but the correlation was inverse, as seen in Figure 8 and Figure 9. There were no significant differences from play condition for either modeled or reported challenge; however, the correlation reveals an inverse relationship. In modeling challenge, we assumed that a player's arousal would increase with challenge; however, upon further examination, this pattern was not always true. Some participants' comments revealed a strategy to attempt to relax when challenged, in order to improve their performance. Obviously, how participants handle challenge in a game is an individual strategy and additional work is required before challenge can be modeled accurately.

We also examined the subjective results from the post-experiment questionnaires. Frequencies of responses for which condition was deemed the most fun, most challenging, and most exciting were tabulated, as were frequencies for the play condition with the maximum modeled fun, challenge, and excitement. For fun, subjective choice and modeled choice were matched for 5 of the 6 (83%) participants; for excitement, subjective choice and modeled choice matched for all 6 (100%) participants. For challenge, only 1 of the 6 (17%) matched. These results corroborate aforementioned results.

Although the trends between conditions are similar for most of the emotions, there are apparent differences in the relative strength of the emotions. Our model represents the

	Computer	*Friend*	*Stranger*	$F_{2,10}$	*Sig.*	*η2*
Boredom	8.5	6.0	6.5	2.7	.118	.35
Challenge	17.3	18.2	22.5	0.55	.594	.10
Excitement	**21.0**	**52.1**	**42.1**	**5.0**	**.032**	**.50**
Frustration	9.7	6.1	7.3	2.4	.145	.32
Fun	**46.7**	**64.2**	**56.9**	**22.1**	**.003**	**.85**

Table 1: Means for modeled emotion, represented as a percentage. There was a significant difference in excitement and fun between play conditions.

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

Table 2: Means for subjective responses on a 5-point scale. A response of "1" corresponded to "low" and "5" to "high". There were no differences between play conditions.

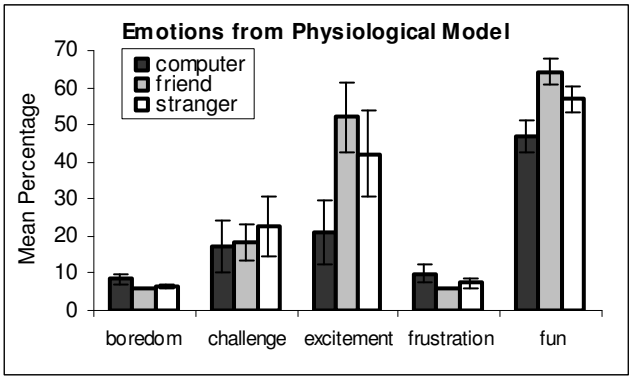

Figure 8: Means (±SE) of modeled emotion, represented as a percentage, separated by play condition.

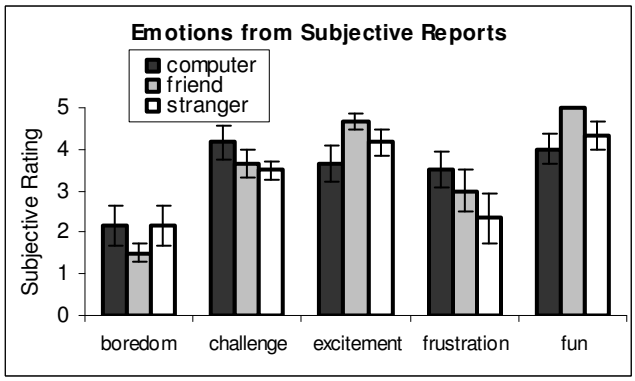

Figure 9: Means (±SE) of the subjective reports on a 5-point scale, separated by play condition.

emotion as a percentage of the possible maximum and minimum, given the available data. Computer games are generally fun, enjoyable experiences. Although a user may be frustrated, and may rate this frustration as fairly high on a 5-point scale, this frustration will be low when compared to the frustration experienced by getting a flat tire on the way to an important appointment, or by trying to contact technical support for a lousy local internet provider. By the same logic, the boredom reported by subjects will be much lower than the boredom experienced during a really boring lecture given by a monotonous professor. We asked participants to agree with the statement "this condition was frustrating". Had we asked them to rate their response as a ratio of how frustrating it was compared to a flat tire on the way to an appointment, we probably would have seen much different subjective results. In contrast, our model takes a global approach to the scaling of emotion, so a user's frustration is given as a percentage of the maximum possible frustration, given the available data. As seen in Figure 8 and Figure 9, boredom, challenge, and frustration are significantly lower for modeled emotion, while fun and excitement are only somewhat lower. This result is expected, since playing a computer game can be quite fun and exciting, but perhaps not as fun and exciting as riding a rollercoaster or attending a rock concert.

MODELED EMOTION: A CONTINUOUS DATA SOURCE

Mean modeled emotion is an objective and quantitative metric for evaluating interactive play technologies that reveals variance between conditions. In addition, modeled emotion from physiological data is very powerful as it can continuously and objectively provide a quantitative metric of user experience *within* a play condition. The mean values shown in Figure 8 are derived from a time series for the five modeled emotions. Figure 10 shows one participant's modeled frustration over time for the three play conditions. The mean values reveal that participant three was most frustrated when playing against the computer, (mean=19.8%), followed by playing against a stranger (mean=13.1%), and playing against a friend (mean=6.5%), but means alone do not tell us whether the tonic level was raised or whether there were more phasic responses.

Figure 10 shows that not only were there more phasic responses (frustrated episodes) when playing against the computer over playing against a friend or stranger, but that these frustrated episodes lasted longer and were greater in amplitude. When playing against a friend, the frustrated episodes were fewer in number, and smaller in amplitude, showing that both tonic level and the number of phasic responses were reduced. Modeled emotion pinpoints moments in time when a user's frustration was changing. This is particularly beneficial when there is no baseline or comparative condition. Researchers and developers can uncover individual moments when a user begins to get stressed, starts having fun, or becomes bored. This information could be used as an evaluative tool, or could be used to dynamically adapt game settings (e.g. difficulty level) to keep players engaged, preventing them from becoming frustrated or bored.

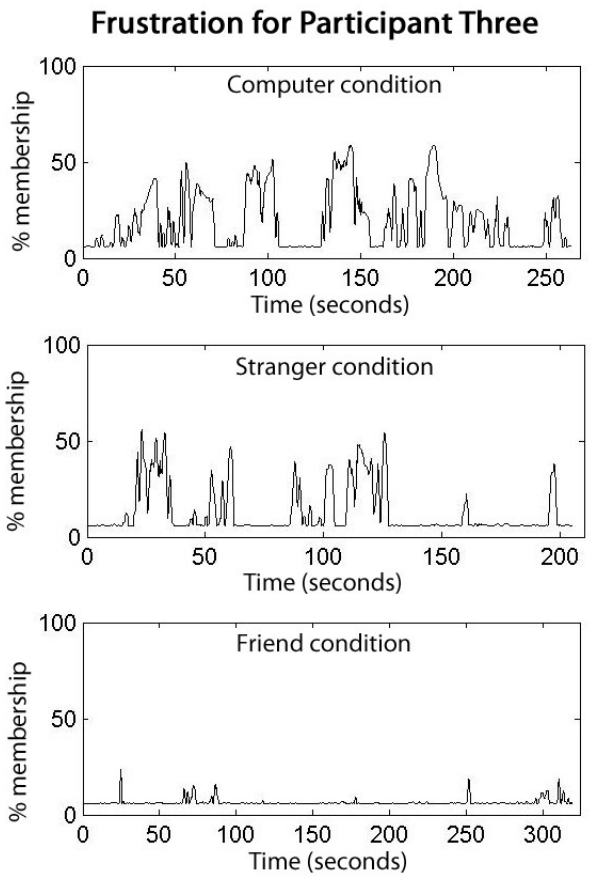

Figure 10: Frustration for one participant in three conditions. Examining the mean output may reveal differences between conditions; however, examining the entire time series reveals how a participant's emotional state changes over time.

SUMMARY

Mean emotion modeled from physiological data provides a metric to fill in the knowledge gap in the objective-quantitative quadrant of evaluating user interaction with entertainment technologies. In addition, the emotion of the user can be viewed over an entire experience, revealing the variance within a condition, not just the variance between conditions. This is especially important for evaluating user experience with entertainment technology, because the success is determined by the *process* of playing, not the *outcome* of playing [23]. The continuous representation of emotion is a powerful evaluative tool that can be easily combined with other evaluative methods, such as video analysis. Given a time series of emotional output, researchers can identify interesting features, such as a sudden increase or decrease in an emotional state, then investigate the corresponding time frame in a video recording. This method would drastically reduce the time required to qualitatively examine video of user interaction with entertainment technologies.

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

FUTURE WORK AND CONCLUSIONS

In addition to integrating the modeled emotion with other evaluation methods, there are other research directions to consider. We developed models for five emotional states that we felt were relevant to interaction with entertainment technology. We would like to consider other relevant emotional states that can be described by arousal and valence, such as disappointment, anger, or schadenfreude. In addition, we would like to see if our method can generalize to interaction with other play technologies, specifically, to study user behaviour in ubiquitous play [2, 16] environments. Once generalized, modeled emotion can be used to dynamically adapt play environments to keep users engaged. When the software determined that players were getting bored, the challenge of the task could increase, or the challenge of the task could decrease if players were becoming overly frustrated. Furthermore, the techniques described in this paper could be adapted to analyze a user's emotional response to productivity software, or other work-related interactive technologies.

We have presented a method of modeling user emotional state when interacting with play technologies. Modeled emotions can be a powerful evaluation approach because they are objective and quantitative (filling a knowledge gap); they account for user emotion; and they present a method of continuous evaluation over an entire condition, revealing process as well as variance. Furthermore, the modeled emotions show the same trends as reported emotions for fun, boredom, and excitement; however, the modeled emotions revealed differences between play conditions, while the differences between the subjective reports failed to reach significance.

We have shown that there is great potential for using physiological metrics to model emotional experience with interactive play technologies.

ACKNOWLEDGMENTS

Thanks to NSERC, NECTAR, EA Sports, and SFU Surrey.

REFERENCES

1. Interactive digital software association: www.idsa.com.
2. Björk, S., Holopainen, J., Ljungstrand, P., and Mandryk, R.L. Introduction to special issue on ubiquitous games. *Personal and Ubiquitous Computing*, *6*,(2002), 358–361.
3. Boucsein, W. *Electrodermal activity.* Plenum Press, New York, 1992.

4. Cacioppo, J.T., Berntson, G.G., Larsen, J.T., Poehlmann, K.M., and Ito, T.A. The psychophysiology of emotion, in *Handbook of emotions*, (M. Lewis and J.M. Haviland-Jones, eds.). The Guilford Press: New York, (2000).
5. Cacioppo, J.T., and Tassinary, L.G. Inferring psychological significance from physiological signals. *American Psychologist*, *45,*1, (1990), 16-28.
6. Chen, D., and Vertegaal, R. Using mental load for managing interruptions in physiologically attentive user interfaces. In *Ext. Abst. CHI '04,* ACM Press (2004), 1513-1516.
7. Desurvire, H., Caplan, M., and Toth, J.A. Using heuristics to evaluate the playability of games. In *Ext. Abst. CHI 2004,* ACM Press (2004), 1509-1512.
8. Ekman, P. Basic emotions, in *Handbook of cognition and emotion*, (T. Dalgleish and M. Power, eds.). John Wiley & Sons, Ltd.: Sussex, (1999).
9. Ekman, P., Levenson, R.W., and Friesen, W.V. Autonomic nervous system activity distinguishes among emotions. *Science*, *221,*4616, (1983), 1208-1210.
10. Fisher, C., and Sanderson, P. (1996). Exploratory data analysis: Exploring continuous observational data. *Interactions, 3* (2), 25-34.
11. Fulton, B., and Medlock, M. Beyond focus groups: Getting more useful feedback from consumers. In *Proc. Game Dev. Conf.,* (2003).
12. Lang, P.J. The emotion probe. *American Psychologist*, *50,*5, (1995), 372-385.
13. Lang, P.J., Greenwald, M.K., Bradley, M.M., and Hamm, A.O. Looking at pictures: Affective, facial, visceral, and behavioural reactions. *Psychophysiology*, *30,*(1993), 261-273.
14. Lazzaro, N. Why we play games: 4 keys to more emotion. In *Proc. Game Dev. Conf.,* (2004).
15. Levenson, R.W. Autonomic nervous system differences among emotions. *American Psychological Society*, *3,*1, (1992), 23-27.
16. Magerkurth, C., Cheok, A.D., Mandryk, R.L., and Nilsen, T. Pervasive games: Bringing computer entertainment back to the real world. *ACM Computers in Entertainment*, *3,*3, (2005), Article 4A.
17. Mandryk, R.L., and Atkins, M.S. (2006). *A fuzzy physiological approach for continuously modeling emotional experience during interaction with computer games*, Technical Report: Simon Fraser University, Burnaby, BC.
18. Mandryk, R.L., and Inkpen, K. Physiological indicators for the evaluation of co-located collaborative play. In *Proc. CSCW 2004,* ACM Press (2004), 102-111.
19. Mandryk, R.L., Inkpen, K., and Calvert, T.W. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour and Information Technology (Special Issue on User Experience)*, *25,*2, (2006), 141-158.
20. Marshall, C., and Rossman, G.B. *Designing qualitative research.* Sage Publications, Thousand Oaks, 1999.
21. Nielsen, J. Evaluating the thinking-aloud technique for use by computer scientists, in *Advances in human-computer interaction*, (H.R. Hartson and D. Hix, eds.). Ablex Publishing Corporation: Norwood, (1992), 69-82.
22. Norman, D.A. (2002). Emotion and design: Attractive things work better. *Interactions, 9* (4), 36-42.
23. Pagulayan, R.J., Keeker, K., Wixon, D., Romero, R., and Fuller, T. User-centered design in games, in *Handbook for human-computer interaction in interactive systems*, (J. Jacko and A. Sears, eds.). Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, (2002), 883-906.
24. Papillo, J.F., and Shapiro, D. The cardiovascular system, in *Principles of psychophysiology: Physical, social, and inferential elements*, (L.G. Tassinary, ed. Cambridge University Press: Cambridge, (1990), 456-512.
25. Partala, T., and Surakka, V. The effects of affective interventions in human-computer interaction. *Interacting with Computers*, *16,*(2004), 295-309.
26. Picard, R.W. *Affective computing.* MIT Press, Cambridge, MA, 1997.
27. Rowe, D.W., Sibert, J., and Irwin, D. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *Proc. CHI '98,* (1998), 480-487.
28. Russell, J.A., Weiss, A., and Mendelsohn, G.A. Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57,*3, (1989), 493-502.
29. Scheirer, J., Fernandez, R., Klein, J., and Picard, R. Frustrating the user on purpose: A step toward building an affective computer. *Interacting with Computers*, *14,*2, (2002), 93-118.
30. Stern, R.M., Ray, W.J., and Quigley, K.S. *Psychophysiological recording.* Oxford University Press, New York, 2001.
31. Sweetsner, P., and Wyeth, P. GameFlow: A model for evaluating player enjoyment in games. *ACM Computers in Entertainment*, *3,*3, (2005), Article 3A.
32. Vicente, K.J., Thornton, D.C., and Moray, N. Spectral analysis of sinus arrhythmia: A measure of mental effort. *Human Factors*, *29,*2, (1987), 171-182.
33. Ward, R.D., and Marsden, P.H. Physiological responses to different web page designs. *International Journal of Human-Computer Studies*, *59,*1/2, (2003), 199-212.
34. Wilson, G.M., and Sasse, M.A. Do users always know what's good for them? Utilizing physiological responses to assess media quality. In *Proc. HCI 2000,* Springer (2000), 327-339.
35. Wilson, G.M., and Sasse, M.A. Investigating the impact of audio degradations on users: Subjective vs. Objective assessment methods. In *Proc. OZCHI 2000,* (2000), 135-142.
36. Winton, W., Putnam, L., and Krauss, R. Facial and autonomic manifestations of the dimensional structure of emotion. *Journal of Experimental Social Psychology*, *20,*(1984), 195-216.