Entertainment Modeling through Physiology in Physical Play

Georgios N. Yannakakis *

Center for Computer Games Research
IT-University of Copenhagen
Rued Langgaards Vej 7, Copenhagen S, DK-2300

John Hallam

Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Campusvej 55, Odense M, DK-5230

Abstract

This paper is an extension of previous work on capturing and modeling the affective state of entertainment ("fun") grounded on children's physiological state during physical gameplay. The goal is to construct, using representative statistics computed from children's physiological signals, an estimator of the degree to which games provided by the playground engage the players. Previous studies have identified the difficulties of isolating elements of physical activity attributed to reported entertainment derived (solely) from heart rate recordings. In the present article, a survey experiment on a larger scale and a physical activity control experiment for surmounting those difficulties are devised. In these experiments, children's Heart Rate (HR), Blood Volume Pulse (BVP) and Skin Conductance (SC) signals, as well as their expressed preferences of how much "fun" particular game variants are, are obtained using games implemented on the Playware physical interactive playground. Given effective data collection, a set of numerical features is computed from these measurements of the child's physiological state. A comprehensive statistical analysis shows that children's reported entertainment preferences correlate well with specific features of the recorded signals. Preference learning techniques combined with feature set selection methods permit the construction of user models that predict reported entertainment preferences given suitable signal features. The most accurate models are obtained through evolving artificial neural networks and are demonstrated and evaluated on a Playware game and a control task requiring physical activity. The best network is able to correctly match expressed preferences in 69.64% of cases on previously unseen data (p-value = 0.0022) and indicates two dissimilar classes of children: those that prefer constantly energetic play of low mental/emotional load; and those that report as fun a dynamic play that involves high mental/emotional load independently of physical effort. The generality of the methodology, its limitations, its usability as a real-time feedback mechanism for entertainment augmentation and as a validation tool are discussed.

Key words: Affective computing, Fun, Entertainment Modeling, Physical Games, Preference learning, Physiology, Heart rate, Blood volume pulse, Skin conductance

---

* Corresponding author
Email addresses: `yannakakis@itu.dk` (Georgios N. Yannakakis), `john@mmmi.sdu.dk` (John Hallam).

1 Introduction

The principal goal in the reported work is to construct children-user models of a class of game-playing experience during physical play in the Playware playground platform. Specifically, the aim is a model that can predict the children's answers to which variants of a given game are more or less "entertaining" (or "fun," which is used synonymously in this paper). The word "fun" is used extensively hereafter since it captures best, in our view, children's notion of the term "entertainment" [1] and is the term used by the children when making their experimental self-reports. This approach is referred to as *Entertainment Modeling*. Entertainment generated by a physical game experience is captured through features extracted from the player's physiological state and feature selection is used for choosing appropriate sets of features that successfully predict expressed entertainment preferences. Game play experiences might very well be identified instead. or as well, through features extracted from user-game interaction. Furthermore, game play behavior could be video recorded and emotions could be recognized by experts or automatically through face gesture detection; however, these approaches are not the focus of this work.

Entertainment is a highly complicated mental state. However, it is correlated with sympathetic arousal [2] and this can be measured, as reported by researchers in the psychophysiological research field [3], using specific physiological signals such as heart rate variability (HRV) and skin conductivity. Although the impact on a subject's physiological state of emotional engagement during computer game playing is well reported in the literature (see [2] among others), there are very few corresponding studies in the physical play domain.

Motivated by the lack of entertainment modeling approaches grounded on a player's physiological state in physical play, the Playware [4] physical interactive game platform has been used for recording Heart Rate (HR) signals of children during play [5]. In that study the complexity of isolating the HR elements of physical activity from expressed entertainment in physical games were outlined in game experiments with 56 child participants. This complexity was handled, in part, through a carefully designed control experiment of physical activity [5].

In the present article entertainment models are constructed using three alternative preference learning techniques (large margin algorithm [6], meta-large margin algorithm and neuro-evolution) applied to statistical features derived from physiological signals measured during play and children's self-report preference data. The output of the constructed models is a real number y such that more enjoyable games receive higher numerical output and functions as an efficient predictor of reported entertainment preferences given suitable specific physiological signal features. Suitable input feature sets are constructed using two alternative feature selection schemes (n best features selection and sequential forward selection), the performances of which are compared. This basic approach of entertainment modeling is applicable to a variety of games, both computer [7] and physical [8,9,5] using features derived from physiological data and/or from the interaction of player and opponent measured through game parameters.

As a sequel to previous work [10,5] a new set of experiments for capturing entertainment preferences through physiology in physical play is presented here. This experiment expands the investigation of the physiological state's relation to entertainment preferences from HR to include Blood Volume Pulse (BVP) and Skin Conductance (SC) signals; employs automatic techniques to identify important features for model construction; and compares preference learning methods as model-building tools. Moreover, the number of child participants is increased to 72 allowing for the creation of more accurate and generic user models. To control for elements of physical activity influencing the physiology of entertainment, an objectively (by human-verification) non-entertaining form of physical activity needs to be tested. For this purpose, a second game experiment, first introduced in [5], is employed, where a control physical activity task with characteristics similar to game activity is compared with game activity by children.

A statistical analysis reveals that features extracted from HR and BVP that correspond to both physical and mental/emotional effort correlate significantly with expressed preferences. Moreover, preference learning attempts on single features indicate that the energy of the high frequency band of HRV (derived from power spectral analysis) constitutes the feature that performs best in predicting expressed preferences on unknown data. This feature, which is suppressed during mental or emotional stress [11,12], is highly anti-correlated to reported entertainment indicating high parasympathetic heart activity on preferred games. This analysis also suggests that collecting physiological signals beyond HR, such as BVP, may provide more meaningful features (e.g. energy of the high frequency band of HRV) for capturing entertainment preferences of children in physical play.

Comparative studies between the two feature selection methods and the three preference learning approaches reveal that evolving Artificial Neural Network (ANN) models combined with sequential forward selection generate the highest accuracy in classifying between preferred and not preferred Playware game variants. These models are trained and validated on game-play data obtained from the first (main) set of experimentation and then are evaluated using unseen data from the second game-play and control experiment set. The results indicate that ANN user models able to predict children's preferred game variants given suitable HR and HRV feature representations can indeed be constructed and that such models not only distinguish game-play from game-like non-entertaining physical activity but also generalize (to some extent) over children's individual preferences.

The paper concludes with a discussion of the limitations of the proposed methodology and of the extent to which it could be applied to other genres of digital entertainment. Its generic use as an efficient baseline for capturing reported entertainment in physical interactive games in real-time is also outlined.

2 Capturing Entertainment through Physiology

Measurements of physiological quantities have been used extensively within the affective computing research area for emotion recognition in children and adults. HR and HRV have been used to effect discrimination between children's exploration, problem-solving and play tasks [13]. Experiments with two-year old children further showed suppression of HRV during exploration, and solution of a puzzle, suggesting that the task demands for these two activities were greater than those during play [14].

Correlations between physiological signals — galvanic skin response (GSR), jaw electromyography, respiration and cardiovascular measures — and reported adult user experiences in computer games have been examined by Mandryk et al. [2]. Statistical analysis yields a significant correlation only between GSR and reported "fun" in video games in that study. In [15], a fuzzy model with rules grounded in psychophysiology theory indicates that high arousal and positive valence (a combination corresponding to "fun" and excitement) is present when HR and GSR are high and the electromyography (EMG) in jaw corresponds to a smiling player. The model is validated against subjects' reported "fun" through correlation-based statistical tests and provides an objective notion of "fun" based on the relation between the data obtained and subjects' expressed preferences.

Working on the same basis as Mandryk et al. [2], Ravaja et al. [16] examined whether the nature of the game opponent influences the physiological state of players. In addition, Hazlett's [17] work focused on the use of facial EMG to distinguish positive and negative emotional valence during interaction with a racing video game.

Rani and colleagues' preliminary experiments [18] are closely related to our work. They demonstrate appropriately adjusting the level of challenge in the game of 'Pong' using physiological signals recorded in real-time and subject's self-reports of their emotional experiences during gameplay. The study, however, is primarily focused on anxiety level detection in real-time and involves a rather limited number of human participants. Physiological state (HR, SC) prediction models have also been proposed for potential entertainment augmentation in computer games [19].

Day-dependence and methodological conditions in capturing and classifying emotions when using physiological signal data raised by Picard et al. [20] are satisfied in the work described in this paper. All experiments described meet three of the five factors for eliciting genuine emotion in the most natural setup (*ibid.*): the experiments took place in a setup close to the *real-world* since children played in their school classroom, our emphasis was on internal *feelings* and subjects were not aware of the purpose of the experiment (*other-purpose*). Note that the study presented here is not focussed on the investigation of the long-term realistic physiology of children with regards to entertainment but rather the construction of a predictor of reported entertainment based on individual physiological signal features.

All of the studies referred to above use physiological measurements for capturing user experiences (e.g. “fun”, engagement or excitement) applied within the computer and edutainment games framework. Motivated by the innovative output of earlier studies on the interplay between HR signal features and reported entertainment in physical play domain [10,5,21] the work reported here is novel in that it examines physiological state (HR, BVP, SC) correlates of reported “fun” in physical activity games, attempts to isolate physiological signal features attributed to reported entertainment in such physically demanding games and proposes a way of constructing a subjective model (a predictor of user preferences) of reported “fun” grounded in statistical features of physiological signal dynamics.

3 Test-bed Physical Games

The Playware [4] prototype playground consists of building blocks (i.e. tangible tiles) that allow for the game designer (e.g. the child) to develop a significant number of different games within the same platform. The overall technological concept of Playware is based on physically implemented computational agents (the tiles) incorporating processing power, communication, input and output. The *Digiwall* [22] and *Age Invaders* [23] mixed-reality systems, the *Scorpiodome* [24] game system, the *STARS* [25] tabletop game and the *Ping-Pongplus* [26] digitally enhanced ping-pong game are platforms closely related to the augmented-reality Playware. See [4,8,5] for further details on the Playware playground.

The “Bug-Smasher” game is used as the test-bed game in the experiments presented here. Bug-Smasher is developed on a 6 × 6 square tile topology. During the game, different ‘bugs’ (colored lights) appear on the game surface and disappear sequentially after a short period of time by turning a tile’s light on and off respectively. A bug’s position is picked randomly according to a predefined level of bug spatial diversity. The child’s goal is to smash as many bugs as possible by stepping on the lighted tiles. Bug-smasher has been used as a test-bed in previous work; further details can be found in [27,8,10,5].

4 Experiment Setup

Following the experimental design proposed in [28,5] for effectively capturing the level of entertainment, the test-bed game under investigation is played in variants. For this purpose, different states (e.g. ‘Low’, ‘High’) of quantitative estimators of qualitative entertainment factors (e.g. challenge, curiosity and fantasy [29]) are used. (The reader may refer to [30,5] for an analysis of quantitative measures of the challenge and curiosity factors in the Bug-Smasher game.) The combination of states/number of entertainment factors generates a pool of dissimilar games for the designer to investigate.

By experimental design (see [7,8] and Fig. 1), each subject plays against k of the n variants of the selected game in all permutations of pairs. k equals 2 and n equals 9 in the main experiment presented in this paper. Thus, C_k^n is the required number of subjects to cover all combinations of k out of n game variants. More specifically, each child plays games in pairs (game A and game B) — differing in the levels/states of one or more of the selected entertainment factors — for a selected time window. Each time a pair of games (‘game pair’) is finished, the child is asked whether the first game was more “fun” than the second game (pairwise preference). Children are not interviewed but are asked to fill in a questionnaire, minimizing the interviewing effects reported in [2]. To minimize any potential order effects we let each child play the aforementioned games in both orders. Statistical analysis of the effect of order of game playing on children’s judgement of entertainment indicates the level of randomness in children’s preferences (see Section 4.1.1 and Section 4.2.1). Randomness is apparent when the child’s expressed preferences are inconsistent for the pair (A, B); i.e. $A \succ B$ and $B \succ A$.

All subjects are given the same instructions by an experimenter who is unaware of the purpose of the experiment (see Fig. 1). No further oral or eye-contact communication takes place during experiment tasks and questionnaire, minimizing experimenter expectancy effects [31]. The playing time window chosen (90 seconds in this paper) is a compromise between effective data collection (long enough subject-game interaction to support a relative judgement) and not overstretching children with excessive periods of energetic physical play.

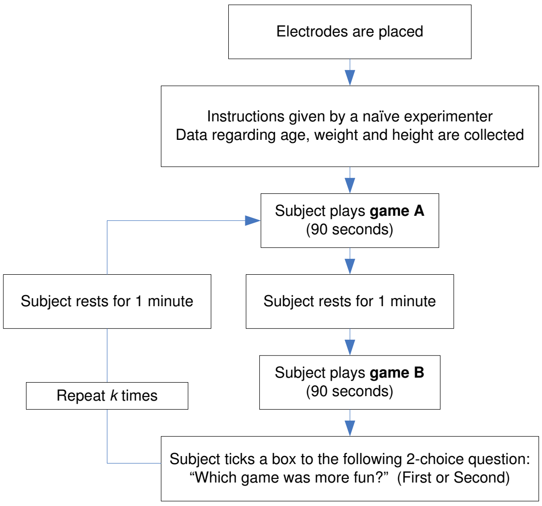


Fig. 1. General phases of the experimental setup followed

Capture of emotions, such as entertainment, is considered, in general, a hard problem mainly because understanding emotion is hard [20]. Capturing reports of playing experiences or emotions is still tough since data obtained embed experimental noise and subjectivity. As previously mentioned, a pairwise preference scheme (2-alternative forced choice: 2-AFC) is used in self-reports of children. 2-AFC offers several advantages for subjective entertainment capture: it minimizes the assumptions made about subjects' notions of "fun" and allows a fair comparison between the answers of different subjects. Since the focus is to construct a model relating reported entertainment preferences to individual playing features that generalizes over the reports of different players, 2-AFC is preferred to a ranking approach [2]. On the other hand, 2-AFC generates "noise" in those cases where the subject has no strong preference to express. The extent to which this "noise" is non-random is detected by the order effect statistics mentioned above.

Heart rate h, blood volume pulse b and skin conductance s are the three signals recorded during gameplay in all experiments presented in this article. These signals are selected because of their popularity among physiological signals used in affective computing studies [2] and their correlation to sympathetic arousal. Several studies have reported the direct correlation between HR and sympathetic arousal (see [3] among others). Moreover, BVP signals can display changes in sympathetic arousal since the sympathetic nervous system controls the size of the blood vessels (vasomotor activity): an increase in the BVP amplitude indicates greater blood flow to the peripheral vessels (e.g. fingertips) and consequently decreased sympathetic arousal [20]. BVP signals are obtained through photoplethysmography which is the process of applying a light sensor to an appendage (e.g. fingertip), and measuring the light that is reflected by the skin. Finally, changes in skin conductance reflect changes in the level of arousal in the sympathetic nervous system. Increased SC indicates heightened sympathetic nervous system arousal [20,2]. To measure SC, a small voltage is applied through two electrodes to the skin (e.g. fingertip) and the skin's current conduction is measured.

For measuring the above-mentioned signals, the Thought Technologies ProComp Infiniti biosensoring system is used, which is housed in a custom-designed belt worn by children with three sensors (two electrodes for SC and one photo sensor for BVP) placed on their fingertips. By using small and accurate commercial apparatus such as ProComp Infiniti in the least intrusive way we attempt to minimize (psychological) experiment effects caused by the presence of recording devices. BVP and SC signals are sampled 256 times per second whereas HR (h) instances are automatically computed by ProComp Infiniti through the inter-beat time intervals of the BVP signal and stored every 5 seconds. Note that in the presented studies subjects played all their assigned games on the same day, mitigating day-dependence effects on their physiology [20]. Cultural differences in the impact of affect on physiology may also be present but are not examined here.

4.1 Main Experiment

Seventy two normal-weighted (based on their body mass index) children whose ages cover a range between 8 and 10 years participated in the main experiment presented here. 9 Bug-Smasher game variants were created by varying two entertainment factors, challenge (bugs' speed) and curiosity (bugs' spatial diversity), with three states each ('Low', 'Average' and 'High'). The 72 children were asked to play a pair of Bug-Smasher variants according to the protocol presented above.

Out of the total number of 288 games played in this experiment, in 230 games (115 game pairs) and 170 games (85 game pairs) the BVP (and HR) and SC signals respectively were properly recorded. In the remaining games, physiological data was lost because of hardware failure. Malfunction of the recording device (electrodes and data transmission) during the game was the main cause of the recording failure. Electrode misplacement due to movement of very energetic children appears to be the most common factor of data loss. While a loss of data up to 40% (SC signal) of the experimental games is substantial, there is no reason to suppose that the hardware failure has any particular bias with respect to experimental hypothesis. The set of signal time series collected from 85 game pairs of the main experiment where all three signals were correctly recorded underlies the analysis presented in this paper

4.1.1 Order Effects

To avoid any order effect of playing on physiology we allow each child one minute to rest in between the two games of each pair. This amount of time is generally adequate for the child's arousal to drop to the level it was just before the first game started, as observed in previous studies [5]. Indicatively, a t-test for means of paired samples demonstrates no significant difference in both the initial HR ($t = -0.412$, p-value $= 0.341$) and the initial SC ($t = -0.912$, p-value $= 0.182$) values between the two games of the pair.

To check whether the order of playing Playware games affects the children's judgement of entertainment, we follow the order testing procedure previously described in [5] which is based on the times that the subject prefers the first or the second game in both pairs. This statistical analysis, introduced in [27], shows that no significant order effect occurs ($r_c = -0.102$, p-value $= 0.224$). The reported insignificant order effect also, in part, demonstrate that effects such as a child's possible preference for the very first game played and the interplay between reported entertainment and familiarity with the game are statistically insignificant.

4.2 Controlled Physical Activity Experiment

In a physical game context, the degree to which a game is entertaining influences the physical engagement of the player, and hence the intensity of the physical activity as well as (possibly) the kind of physical activity. To control for the former, we designed an additional experiment where the physical activity control is achieved through a non-entertaining variant of the Bug-Smasher game named the "Stomping game." Experiments with this game were first introduced in [5] following a good suggestion from an anonymous reviewer.

The Stomping game is as follows. Children are asked to stomp on a different one of four constantly lighted tiles of different color each time they hear a sound coming from the game platform. The four tiles are placed at the corners of a 3×3 square in the center of the 6×6 platform; two tiles equals the average distance between bugs appearing in the Bug-Smasher game. The sound determining the frequency of child-game interaction occurred at a rate equal to the average of the bug appearance rates of the two different levels of challenge used in the Bug-Smasher game.

This control game is designed on the basis of Malone's studies, in which features of the studied game are subtracted and the effect such changes have on children's entertainment is investigated [29]. The Stomping game eliminates putative entertaining features from Bug-Smasher while retaining a similar requirement for physical activity, making it an appropriate control. Crucially, it lacks an essential element for successful game design, which is the provision of an apparent goal for the user [29]. Moreover, the curiosity entertainment feature is minimal since the game is completely predictable and interaction is absent in that the playground does not react to the child's actions (i.e., bugs are not smashed — turn red and disappear — when pressed as they do in Bug-Smasher). Given these changes, one would not expect the Stomping game to be entertaining for children; and children's self-reports confirm that in the majority of cases this is so.

For the control experimental protocol, we asked 18 naïve normal-weighted children (9 boys and 9 girls) aged 8 to 10 years to play 5 games each on the Playware platform. The set of 5 games played comprised 4 games of Bug-Smasher, in two pairs, and the physical activity control game referred to above. As in the main experiment, two game variants with differing levels of challenge and curiosity were played in both orders, giving four Bug-Smasher variant games plus the control game.

All details regarding the protocol of the experiment follow the principles of the experimental setup described above. To minimize order effects involving the control game, the Stomping game is placed either first, third or last in the sequence of five games played with equal probability. Additionally, children complete the comparison questionnaire after games 2, 3, 4 and 5, resulting in four fun comparisons (expressed preferences) between games 1–2, 2–3, 3–4, and 4–5, for each child to report. That provides a total of 72 (18 children times 4 comparisons) "fun" comparisons including 10 "fun" comparisons between identical game variants which are not further considered, 24 "fun" comparisons between a Bug-Smasher game variant and the Stomping game, and 38 "fun" comparisons between different variants of the Bug-Smasher game. The game pairs under investigation are thus 62 in total. In the "fun" comparisons between the Bug-Smasher game and the Stomping game, children expressed a preference for the Bug-Smasher game in 22 out of 24 cases — confirming the expectation that the Stomping game is a comparatively non-entertaining physical game. All three physiological signals were recorded correctly in 56 out of 62 game pairs of the control experiment, reflecting a loss of approximately 10% of data due to hardware failure of the biosensor system.

4.2.1 Order Effects

Following the protocol of the main experiment, we allow each child one minute to rest between two sequential games in order to avoid any order effect of playing on physiological signals. Paired sample t-tests for means of the initial HR ($t = 0.689$, p-value $= 0.493$) and SC ($t = 0.922$, p-value $= 0.179$) values show that the child's arousal at the start of the two games is well matched.

To check whether the order of playing the games of control experiment affects the children's expressed entertainment preferences, we use the same procedure presented in Section 4.1.1. However, when the stomping game is placed either first or last in the game sequence we are not considering the Stomping-Bug smasher (first or last) game pair whereas when the control game is placed in the middle the procedure is based on counting the times that the subject prefers the first or the second game in the Stomping-Bug smasher game pair only. The obtained r_c value is -0.166 with corresponding p-value is 0.2025, demonstrating that the order of play does not significantly affect children's preferences.

5 Features Extracted

While no transform methodology is applied for the HR signal, the BVP and SC raw signals of both the main and the control experiment are noise-filtered via truncation of their discrete Fourier transform (DFT). A spectral threshold of 20% of the DFT maximum amplitude is used for the experiments presented here. Measurement units for HR, BVP and SC are respectively heart beats per minute (bpm), percent of blood vessel pressure (BVP is a relative measure) and micro-Siemens (μS), an SI measure of conductance (inverse of megohm). Given the noise-filtered signals, the following features are extracted for each signal type:

HR Features extracted from the HR signal are presented in Table 1. In addition to those features, three different regression models were used to fit (least square fitting) the HR signal: linear, quadratic and exponential. The additional features were the parameters of the three regression models mentioned above.

BVP Table 2 presents the features extracted from the BVP signal. Moreover, given the inter-beat (RR) time intervals of the BVP signal the following Heart Rate Variability (HRV) parameters were computed:

- HRV - time domain: the standard deviation of RR intervals $\sigma\{RR\}$, the fraction of RR intervals that differ by more than 50 msec from the previous RR interval $pRR50$ and the root-mean-square of successive differences of RR intervals RMS_{RR} [12].
- HRV - frequency domain: the frequency band energy values derived from power spectra obtained using discrete Fourier transformation; energy values are computed as the integral of the power of each of the following four frequency bands (see [12,33] among others): Ultra Low Frequency (ULF) band: $[0.0, 0.0033]$ Hz; Very Low Frequency (VLF) band: $(0.0033, 0.04]$ Hz; Low Frequency (LF) band: $(0.04, 0.15]$ Hz and High Frequency (HF) band: $(0.15, 0.4]$ Hz.

SC All extracted features used for the HR signal. Additional features include

Table 1
Features extracted from the HR signal.

Feature	Description
$E\{h\}$	Average HR.
sigma$\{h\}$	Standard deviation of HR.
min$\{h\}$	Minimum HR.
max$\{h\}$	Maximum HR.
D_h	Difference between maximum and minimum HR (D_h =max$\{h\}-$min$\{h\}$).
R_h	Correlation coefficient between HR recordings and the time t at which data were recorded.
ρ_1^h	Autocorrelation (lag equals 1) of the HR signal.
h_{in}	Initial HR recording.
h_{last}	Last HR recording.
t^h_{max}	Time when maximum HR occurred.
t^h_{min}	Time when minimum HR occurred.
$t^h_{max} - t^h_{min}$	The time difference between maximum and minimum HR.
$ApEn_h$	Approximate entropy [32] of the signal which quantifies the unpredictability of fluctuations in the HR time series (see [5] for further details on $ApEn$).

Table 2
Features extracted from the BVP signal.

Feature	Description
$E\{b\}$	Average BVP.
sigma$\{b\}$	Standard deviation of BVP.
min$\{b\}$	Minimum BVP.
max$\{b\}$	Maximum BVP.
$E\{IBAmp\}$	Average inter-beat amplitude.
$\delta^b_{\lvert 1\rvert}$	Mean of the absolute values of the first differences of the BVP signal [20].
$\delta^b_{\lvert 2\rvert}$	Mean of the absolute values of the second differences of the BVP signal [20].

the mean of the first and second differences of the raw SC (δ^s_1 and δ^s_1 respectively) and the mean of the absolute values of the first and second differences of the SC signal ($\delta^s_{\lvert 1\rvert}$ and $\delta^s_{\lvert 2\rvert}$ respectively).

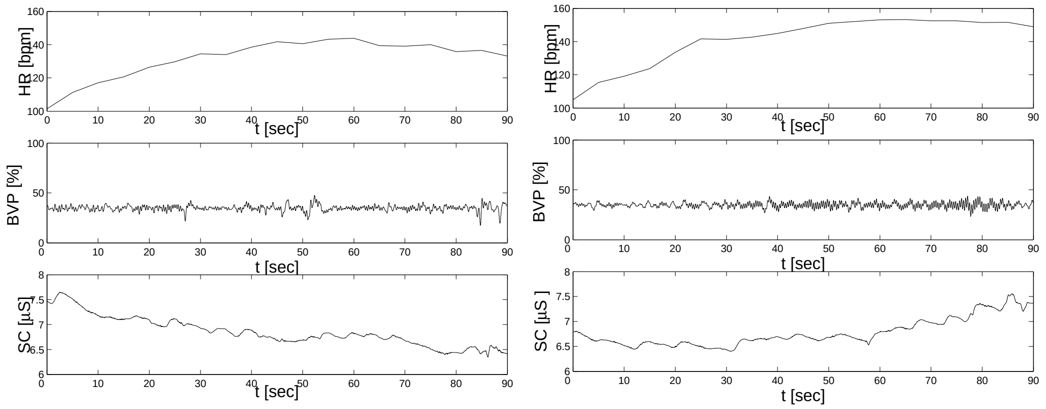
Other statistics are also possible; however, the choice of the specific statistical features was made in order to cover a decent amount of the HR, BVP and SC signal dynamics proposed in the majority of previous studies in the field [20,12,5]. Fig. 2 illustrates an example of the physiological signal dynamics recorded from two different children playing a pair of Bug-Smasher games. Fig. 2(b) and Fig. 2(d) correspond to the game selected by the child as more entertaining of the two. For reasons of space, we present the physiology dynamics of only two nevertheless representative game pairs. Note that the qualitative features of the signals are similar for all children that participated in the main and control experiments. For the HR signal, the general observation is an initial rapid increase of HR during the first seconds of the game followed by a stable, but noisy, condition of high HR values. Moreover, the general trend is that HR is heightened in preferred games compared to non-preferred games (see Fig. 2). Preferred games appear to generate decreased BVP amplitude and increased SC indicating possible heightened sympathetic arousal (see Fig. 2(b) and Fig. 2(d)).

Even though there appear to be correlations between signal dynamics and entertainment preferences a statistical analysis of the extracted features and a preference learning methodology will better determine the extent to which each statistical feature has an impact on reported entertainment.

5.1 Main Data Set Statistical Analysis

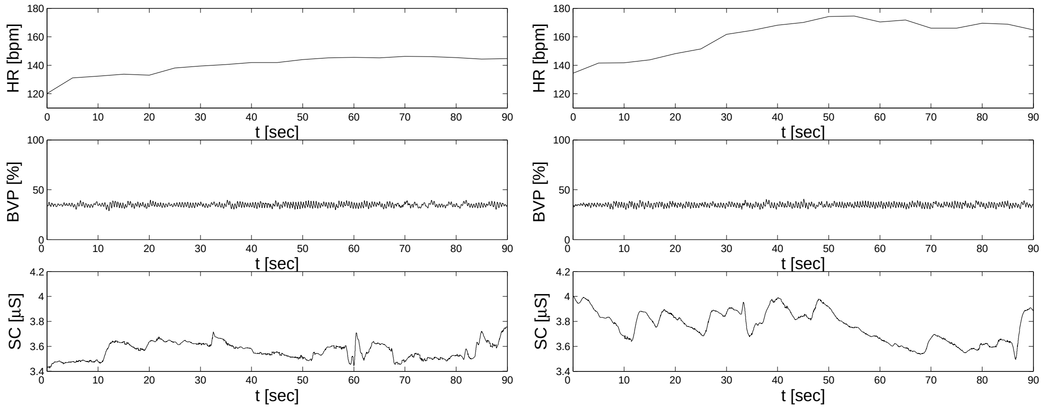
This section presents an analysis for exploring statistically significant correlations between children's expressed preferences and recorded physiological signal features in the main experiment. The above-mentioned correlation coefficients are obtained through $c(\overrightarrow{z}) = \sum_{i=1}^{N_s}\{z_i/N_s\}$, where N_s is the total number of game pairs where physiological signals were properly recorded ($N_s = 115$ for HR and BVP and $N_s = 85$ for SC) and $z_i = 1$, if the subject chooses as the more entertaining game the one with the larger value of the examined feature and $z_i = -1$, if the subject chooses the other game in the game pair i.

Within the HR signal extracted features, significant correlations are observed between average and maximum HR and reported entertainment preferences (see Table 3). These effects are consistent with the significant correlations of both $E\{h\}$ and $\max\{h\}$ on physiological data obtained from previous experiments using the Bug-Smasher game [10,5]. Within the class of features extracted from the BVP signal, significant effects are observed on the mean of the absolute values of both the first and the second differences of the raw signal ($\delta^b_{|1|}$, $\delta^b_{|2|}$) and on the energy of the HF band. On the contrary, no significant effect appears in the class of SC features.



(a) Subject no. 11: Non-preferred (non-"fun") game of the pair

(b) Subject no. 11: Preferred ("fun") game of the pair



(c) Subject no. 19: Non-preferred (non-"fun") game of the pair

(d) Subject no. 19: Preferred ("fun") game of the pair

Fig. 2. Example of the HR, BVP and SC signals obtained through the main experiment in two pairs of Bug-Smasher game: A subject (no. 11) plays a game of *Low* challenge and *Average* curiosity (a) and then a game of *High* challenge and *Average* curiosity (b). Another subject (no. 19) plays a game of *Average* challenge and *Low* curiosity (d) and then a game of *High* challenge and *Low* curiosity (c).

Table 3

Correlation coefficients between reported entertainment and individual physiological features. For reasons of space, the three highest absolute correlation coefficient values for each physiological signal type are ranked and presented here. γ_s is the parameter of the quadratic regression ($s_Q(t) = \beta_s t^2 + \gamma_s t + \epsilon$) on the SC signal which quantifies the rotation angle with respect to the x-axis of the quadratic curve. Statistically significant effects appear in bold.

HR	$c(\overrightarrow{z})$	BVP	$c(\overrightarrow{z})$	SC	$c(\overrightarrow{z})$
$E\{h\}$	**0.224**	$\delta^b_{\lvert 1 \rvert}$	**0.216**	$E\{s\}$	0.167
$\max\{h\}$	**0.209**	$\delta^b_{\lvert 2 \rvert}$	**0.216**	γ_s	0.119
$\min\{h\}$	0.179	HF	**−0.216**	$\sigma\{s\}$	−0.119

Obtained effects demonstrate that the higher the $\delta^b_{|1|}$ and $\delta^b_{|2|}$ values, the steeper the BVP signal and the higher the expressed "fun" preferences of children. Moreover, the lower the energy of the HRV HF band — which is driven by respiration and appears to derive mainly from vagal activity [12] — the more children appear to be entertained. Specifically, the energy of the HF range, representing quicker changes in HR, is primarily due to parasympathetic activity of the heart which is decreased during mental or stress load [11,12]. This analysis which is introduced in [21] suggests the conclusion that high mental or stress load appear to be the main factors that guide a child to prefer a game variant more than another.

5.2 Control Data Set Statistical Analysis

By following the statistical analysis presented in Section 5.1 the correlation coefficients between children's entertainment preferences and physiological signal features are obtained for the control data set. Here, N_s equals 56 for all three signal types. As seen in Table 4, average, maximum and approximate entropy of the HR signal demonstrate significant correlations with reported entertainment preferences. HF energy, the root mean square of successive differences (RMS_{RR}) and the standard deviation ($\sigma\{RR\}$) of the inter-beat (RR) intervals are the respective significantly correlated features derived from the BVP signal.

These effects are, in part, consistent with the statistical correlations obtained from the main data set which demonstrates the generality of the effect of specific features ($E\{h\}$, $\max\{h\}$, HF) to reported entertainment over different set of experiments. (Note that, the significant effect of $E\{h\}$ has already been observed in experiments with dissimilar Playware game test-beds in previous studies [10,5].) Moreover, it appears that this generality extends to objectively non-entertaining physical games (i.e. Stomping game).

The obtained statistically significant effects assume a linear relation between the respective features and reported entertainment which may (or may not) provide insight into the appropriate set of features on which to build a successful non-linear model of reported entertainment using preference learning. However, no safe conclusion can be derived for the appropriate feature subset before the proposed machine learning methodology is applied (see Section 6).

Table 4
Correlation coefficients between reported entertainment and individual physiological features of control data set. For reasons of space, the three highest absolute correlation coefficient values for each physiological signal type are ranked and presented here. Statistically significant effects appear in bold.

HR	$c(\overrightarrow{z})$	BVP	$c(\overrightarrow{z})$	SC	$c(\overrightarrow{z})$
$E\{h\}$	**0.393**	HF	**-0.285**	$E\{s\}$	0.178
$\max\{h\}$	**0.357**	RMS_{RR}	**-0.285**	$\sigma\{s\}$	-0.142
$ApEn_h$	**-0.25**	$\sigma\{RR\}$	**-0.25**	t^s_{max}	-0.142

6 Machine Learning

The proposed approach to entertainment modeling is based on selecting a (constrained) minimal subset of individual features and constructing a quantitative user model that predicts the subject's reported entertainment preferences. The assumption is that the entertainment value y of a given game, which models the subject's internal response to playing the game, that is, how much "fun" it is, is an unknown function of individual features which a machine learning mechanism can learn. The subject's expressed preferences constrain but do not specify the values of y for individual games but we assume that the subject's expressed preferences are consistent.

Constraint satisfaction algorithms cannot solve the problem since the variable y under the constraint $y_A > y_B$ for any two given games A and B has no specific domain values. Likewise, any machine learning which is based on learning a target output is inapplicable since target outputs are unknown. By the use of a ranking approach numerical values for the y variable could be made available; however, ranking is an undesired method for the self-report design of comparative "fun" analysis for the disadvantages mentioned earlier. Preference learning [34] is the only applicable type of machine learning for this constrained classification problem. There are several techniques that learn from a set of pairwise preferences such as algorithms based on support vector machines [6], gaussian processes [35] and evolving ANNs. The three preference learning mechanisms used and compared in this study are presented in following subsections.

6.1 Large Margin Algorithm

The Large Margin Algorithm (LMA) introduced in [6] is based on fundamental theory of Support Vector Machines (SVMs) and constitutes the baseline linear preference learning approach to our problem. LMA has shown successful applications in routing problems where, among many approaches, it even outperforms evolving ANNs [6]. In this application, this algorithm is used to investigate subjective entertainment preference functions (y) which are linear combinations of physiological signal features f, i.e. $y(f) = f \cdot w$. The vector $w = (w_1, w_2, \ldots, w_n)$ represents the positive weight variables (i.e. the linear classifier) under optimization attributed to the n features investigated.

The goal is to obtain $y(f_d^p) > y(f_d^{np})$ meaning that the child prefers a game variant with the feature f_d^p to a game with the feature f_d^{np} for each pairwise preference comparison d. If $f_d = f_d^{np} - f_d^p, d = 1, \ldots, m$ — where m is the number of pairwise comparisons — the classifier with large margin can be obtained by solving the following linear programming problem using the Simplex algorithm [36].

$$\begin{aligned} &\text{minimize} \sum_{j=1}^{n} w_j \\ &\text{subject to } w \cdot f_d \geq 1, d = 1, \ldots, m \qquad (1) \\ &\qquad\qquad w_j \geq 0, j = 1, \ldots, n \qquad (2) \end{aligned}$$

6.2 Meta-LMA

This is an algorithm inspired by the LMA algorithm sharing the same goal ($y(f_d^p) > y(f_d^{np})$) and the principal assumption that the subjective entertainment preference function (y) is a linear combination of physiological signal features. According to the meta-LMA algorithm the weight vector $w = (w_1, w_2, \ldots, w_n)$ of n selected features is adjusted to solve the following linear programming problem:

$$\text{maximize } \frac{1}{m} \sum_{d=1}^{m} g(y(f_d), \epsilon) \qquad (3)$$

$$\text{subject to } y(f_d) \geq \delta, d = 1, \ldots, m \qquad (4)$$

where $y(f_d) = y(f_d^p) - y(f_d^{np})$, δ is 0.05 in all experiments presented here, $g(y(f_d), \epsilon) = 1/(1 + e^{-\epsilon y(f_d)})$ is the sigmoid function and $\epsilon = 30$ if $y(f_d) > 0$ and $\epsilon = 5$ if $y(f_d) < 0$. Both the sigmoidal shape of the subjective function and its selected ϵ values are inspired by its successful application as a fitness function in neuro-evolution preference learning problems on Playware test-bed games [27,5].

6.3 Evolving ANNs

Given the high level of subjectivity of human preferences and the highly noisy nature of input data, we believe that more complex non-linear functions such as ANNs might serve our purposes better. Thus, feedforward multilayered Neural Networks for learning the relation between the selected player features (ANN inputs) and the "entertainment value" (ANN output) of a game are used in the experiments presented here. Since there are no prescribed target outputs for the learning problem (i.e. no differentiable output error function), ANN training algorithms such as back-propagation are inapplicable. Learning is achieved through artificial evolution. Details of the neuro-evolution mechanism used can be found in [9,5].

All three preference learning approaches are trained and validated on data obtained from the main experiment (main data set) exactly as described in the foregoing text. The best approach is then evaluated using unseen data from the physical activity control experiment, to determine the extent to which the constructed user model generalizes. Data from this experiment is referred to as "control data set" in the sequel.

7 Feature Selection

The quality of the predictive model constructed by the preference learning schemes outlined above depends critically on the set of input data features chosen. However, it is not possible to determine a priori the suitability of any given feature for the final model. Therefore, we use automatic feature set selection algorithms to explore the space of possible input feature sets, searching for sets that generate the highest discrimination between preferred and non-preferred games. Using the signal features described above (see Section 5), *n Best Features Selection* (nBest) and *Sequential Forward Selection* (SFS) are applied and compared. nBest Selection ranks the individual signal features used individually in order of model performance; the chosen feature set of size n is then the first n features in this ranking. The SFS method, by contrast, is a bottom-up search procedure where one feature is added at a time to the current feature set. The feature to be added is selected from the subset of the remaining features such that the new feature set generates the maximum value of the performance function over all candidate features for addition [37]. Note that both methods are incomplete. Neither is guaranteed to find the optimal feature set since neither searches all possible combinations (they are each a variant of hill-climbing).

The SFS method is used since it has been successfully applied to a wide variety of feature selection problems, yielding high performance values with minimal feature subsets: see [38], for example, for further discussion and for an application to the classification problem of process identification in resistance spot welding. On the other hand, the nBest method is used for comparative purposes, being the most popular technique for feature selection. More advanced methods such as Sequential Floating Forward Search (SFFS) and Fisher Projection (FP) [20] could be used in future experiments and results could be compared to the existing studies. The feature selection procedure followed here evaluates the usability of each one of the features available and obtains the minimal feature subset approximation to the feature subset that performs best in the classification between preferred games and non-preferred games.

Feature selection algorithms select the input feature set under investigation for each learning mechanism; in the case of the evolved ANN the selected features define its input vector. To evaluate the performance of each input feature subset, the available data is randomly divided into thirds and training and validation data sets consisting of 2/3 and 1/3 of the data respectively are assembled. The performance of each user model is measured through the average classification accuracy of the model in three independent runs using the leave-one-out cross-validation technique on the three possible independent training and validation data sets. Since we are interested in the minimal feature subset that yields the highest performance we terminate the feature selection procedure (nBest or SFS) when an added feature yields equal or lower validation performance to the performance obtained without it. (It is for this reason that the search is incomplete — one might obtain better performance by deleting a feature at this point, for instance.)

8 Best Feature Selection

Given the 85 pairs of preferred/non-preferred game comparisons of the main data set, all three preference learning approaches are applied (see Section 6). The data is partitioned (randomly) into 3 groups which are used as 2/3 training and 1/3 validation data subsets with the leave-one-out cross-validation technique to obtain the average classification performance of each approach. Regarding the minimization of evolved ANN size, it was determined that ANN architectures with 10 hidden neurons, are capable of successfully obtaining solutions of high fitness. This was determined by considering the performance of ANN architectures with up to two hidden layers containing up to 30 hidden neurons each.

As observed from Table 5, there is some consistency between features linearly related to reported entertainment and features that predict entertainment preferences based on a non-linear function (ANN). More specifically, all five features that correlate highly with reported entertainment (see Table 3) appear in Table 5. Moreover, given the best and average performance of the eight highest performing features, the non-linear ANN function proves advantageous over linear LMA and meta-LMA in single feature experiments. The best feature for all mechanisms is the HF energy with a best performance of 66.67% (average of 71.43%, 67.86%, 60.71%) obtained through the evolving ANN approach. Given that the average performance of 30 randomly generated ANNs (10 for each validation set) is 48.80%, the P for the HF energy performance to occur is 0.043.

The single feature experiments also suggest that collecting physiological signals beyond HR, such as BVP, may provide more meaningful features (e.g. HF, $\delta^b_{|1|}$, $\delta^b_{|2|}$) for capturing entertainment preferences of children in physical play.

Table 5
The eight highest performing features for each learning mechanism ranked by cross-validation performance (P) from left to right.

LMA	HF	LF	$\sigma\{RR\}$	RMS_{RR}	t^h_{max}	VL	max$\{b\}$	$\sigma\{s\}$
	64.29	63.10	60.71	58.33	58.33	57.14	55.95	55.95
Meta-LMA	HF	LF	$\delta^b_{\|1\|}$	$\delta^b_{\|2\|}$	max$\{h\}$	$\sigma\{RR\}$	h_{in}	$E\{h\}$
	64.29	63.10	61.90	61.90	60.71	60.71	60.71	59.52
ANN	HF	min$\{b\}$	LF	$\delta^b_{\|1\|}$	$\delta^b_{\|2\|}$	max$\{h\}$	$\sigma\{RR\}$	h_{in}
	66.67	63.10	63.10	63.10	61.90	60.71	60.71	60.71

9 More Features

The initial feature subset for all three preference learning approaches includes the feature that performs best in the single feature experiment (HF — cross-validation performance of 66.67%). By applying the SFS method for each learning approach we obtain cross-validation performances presented in Table 6. As expected, results show the advantage of non-linear (ANN) over linear (LMA, meta-LMA) learning approaches for our preference learning case-study. Even though the LMA method compared to neuro-evolution has generated higher performing solutions in specific problems (e.g. routing [6]), the opposite occurs in our study. More specifically, the best cross-validation performance (79.76%; average of 75.00%, 75.00% and 89.29%) is achieved through evolved ANNs with the feature subset $\{\text{HF}, E\{h\}, \sigma\{RR\}\}$ while adding more features to the subset does not yield significantly higher performance (see bottom row of Table 6). On the other hand, the best performances that LMA and meta-LMA achieve are significantly lower; 67.86 and 70.24 respectively. Those mechanisms' generated performance can be used as a baseline for comparison to the best evolved ANN solution, which indicates that the non-linear combination of HF, $E\{h\}$ and $\sigma\{RR\}$ is necessary for a successful predictor of reported entertainment for Playware games.

Comparing feature selection methods within the evolving ANN approach, the SFS method generates feature subsets that yield higher validation performance than feature subsets generated by nBest, as presented in Table 7. The same effect occurs for all three preference learning mechanisms, indicating the benefits of searching with SFS for appropriate feature sets; however, results from nBest are not presented for LMA and meta-LMA for reasons of space.

Obtained best performance (79.76%) appears to be rather low and shows the difficulty in distinguishing physiological signals between games in terms of the reported preferences of entertainment. However, the reported complexity of classifying emotions through physiological state [20], the augmented signal noise recorded during physical play and the binomial-distributed probability of this performance to occur at random (**0.0004**) suggest that the evolved ANNs are successful predictors of children's reported entertainment preferences based on features extracted from physiological state.

Table 6
Classification accuracy (%) of random network, LMA, meta-LMA and ANN approaches by the use of the SFS feature selection method. The highest performance of each approach obtained through a minimal feature subset appears in bold. Random network performance is the average performance of thirty random weight value initializations of the ANN (ten for each validation data set). The random network's input vector consists of the best feature subset that generates the highest cross-validation performance (i.e. $\{\text{HF}, E\{h\}, \sigma\{RR\}\}$).

Random	LMA		Meta-LMA		ANN	
P	Features	P	Features	P	Features	P
50.59	HF	64.29	HF	64.29	HF	66.67
	$\sigma\{s\}$	65.48	$\sigma\{s\}$	**70.24**	$E\{h\}$	71.43
	s_{last}	**67.86**	s_{last}	70.24	$\sigma\{RR\}$	**79.76**
	$\max\{b\}$	67.86	$\max\{b\}$	70.24	t^h_{min}	75.00

Multiple feature experiments confirm the hypothesis that physiology beyond HR may provide features (e.g. HF, $\sigma\{RR\}$) for capturing entertainment preferences more accurately. The best performance obtained is better than reported performance of respective studies grounded solely on HR signals (76.00%) [5]. The $E\{h\}$ feature was included in the highest performing feature subset of that study demonstrating the potential of average HR for entertainment modeling over dissimilar experiments. The addition of HRV statistical features such as HF and $\sigma\{RR\}$ in this experiment results in increased accuracy of the ANN model. Note that SC extracted features, as in single feature experiments, are absent from the highest performing feature subsets.

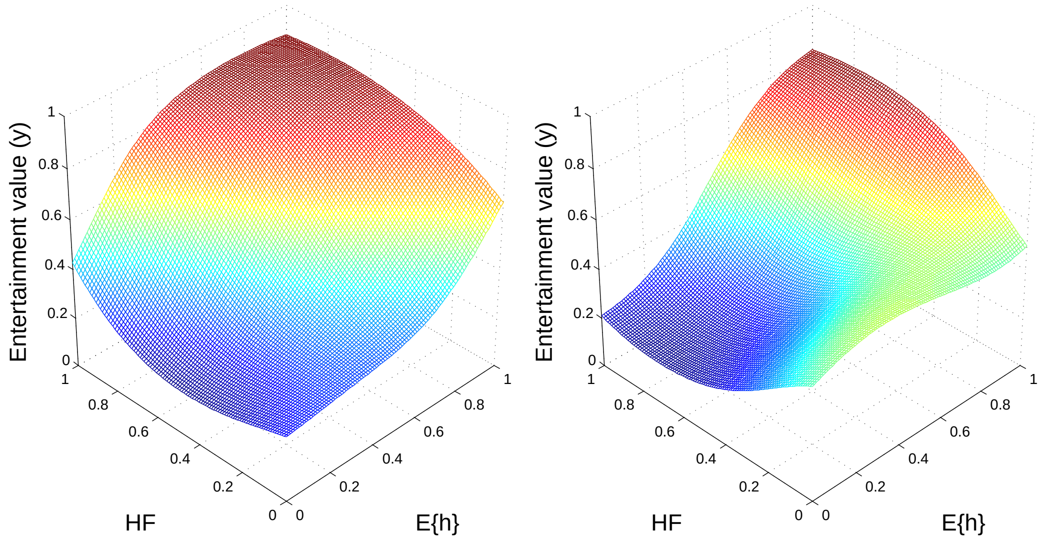
Table 7
Classification accuracy (%) of nBest and SFS feature selection methods applied on the evolving ANN approach. The highest performance obtained through a minimal feature subset appears in bold.

nBest		SFS	
Feature subset	P	Feature subset	P
{HF}	66.67	{HF}	66.67
{HF,min$\{b\}$}	57.14	$\{\text{HF}, E\{h\}\}$	71.43
{HF,min$\{b\}$,LF}	70.24	$\{\text{HF}, E\{h\}, \sigma\{RR\}\}$	**79.76**
{HF,min$\{b\}$,LF, $\delta^{b}_{\lvert 1\rvert}$}	66.67	$\{\text{HF}, E\{h\}, \sigma\{RR\}, t^{h}_{min}\}$	75.00

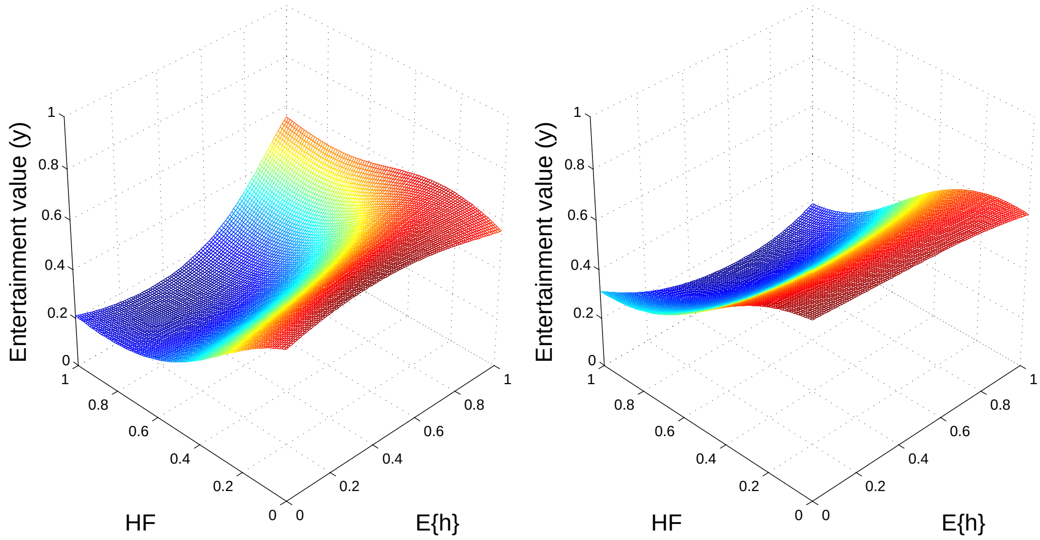
9.1 *Evolved ANN:* $\{HF, E\{h\}, \sigma\{RR\}\}$ *Feature Subset*

A more detailed analysis of the evolved ANN model that yields the best classification accuracy is presented here. Given the $\{\text{HF}, E\{h\}, \sigma\{RR\}\}$ feature subset as inputs, the evolved ANNs correctly match 73.81% (average of the three training trials; $\sigma = 2.06\%$) of children's answers on entertainment. Low training performances are due to the early stopping mechanism included in the genetic algorithm to combat overfitting as described in [5]. The function between HF, $E\{h\}$, $\sigma\{RR\}$ and the game's predicted entertainment value (y) given by the highest performing ANN found is illustrated in Fig. 3. As in results presented in [9,5], all three fittest ANNs generated, each trained on different sets comprising 2/3 of total data, exhibit similar qualitative features of the surface illustrated in Fig. 3.

The general trend appearing in Fig. 3 is that there are two classes of children, based on their physiological state, having dissimilar preferences of entertainment. First, when homogeneity of the inter-beat intervals is high (low $\sigma\{RR\}$ values), the games preferred are those that generate a combination of high $E\{h\}$ and HF values (see Fig. 3(a)). This corresponds to a constantly (low $\sigma\{RR\}$) energetic (high $E\{h\}$) play of low emotional or mental load (high HF) which appears to be entertaining for this class of children. On the other hand, the lower the inter-beat interval uniformity becomes, the more the cases where high entertainment values are generated through low HF energy values (see Fig. 3(b) and Fig. 3(c)). This observation becomes more apparent when $\sigma\{RR\} = 1.0$ (see Fig. 3(d)) where the highest entertainment values correspond to very low HF values (HF= 0.0) independently of average HR. These physiological indices are derived from a second class of children's play where dynamic (high $\sigma\{RR\}$) play that involves high mental or emotional load (low HF) is reported as fun independently of physical effort ($E\{h\}$).



(a) $\sigma\{RR\} = 0.0$ (b) $\sigma\{RR\} = 0.33$



(c) $\sigma\{RR\} = 0.66$ (d) $\sigma\{RR\} = 1.0$

Fig. 3. Evolved ANN that yields the best classification accuracy on unknown data (89.29%): ANN output y (entertainment value) with regards to HF and $E\{h\}$ for four values of $\sigma\{RR\}$ (0.0, 0.33, 0.66, 1.0).

9.2 Validation on Control Data Set

To investigate the extent to which the predictive model of entertainment preference computed using the data from the main experiment generalizes to new experimental data, the best performing evolved ANNs presented in Section 9.1 are presented with and evaluated on the unseen data from the physical activity control experiment.

Table 8 shows the average total classification accuracy (fourth column) and the sub-classification performance for the comparisons between the Bug-Smasher game played and the Stomping game (second column) as well between the Bug-Smasher game chosen as more entertaining and the Bug-Smasher game chosen as less entertaining (third column) of all three evolved ANN.

The performance obtained equals 69.64%, which appears rather low compared to 79.76% of correct matching on the validation data of the main experiment. However, given the reported complexity of the task [20] and the binomial-distributed probability of this performance to occur at random (**<u>0.0022</u>**), the ANN model proves a quite effective, robust and generic predictor of children's reported entertainment preferences based on their physiological state in physical play. The average performance of 10 ANNs identical in structure to the evolved ones, but with random weights, is given for comparison.

Table 8

Evolved ANNs (feature subset: $\{\mathrm{HF}, E\{b\}, \sigma\{RR\}\}$) trained on main data set: Classification accuracy (%) on unseen control data set. The average classification performance of 10 networks with random weights is given for comparison (random network).

	Bug-Smasher game vs. Stomping game	Preferred game vs. non-preferred game	Total
Random network	52.44	50.00	50.99
$\{\mathrm{HF}, E\{b\}, \sigma\{RR\}\}$	75.75	65.68	69.64

10 Conclusions and Discussion

This paper explored the interplay between physiological signals and children's entertainment preferences in physical play. More specifically, the quantitative impact of children's reported entertainment on HR, BVP and SC signal statistics was investigated through an action game (Bug-Smasher) developed on the Playware playground. The statistical effects obtained from the main game experiment presented here provided some first insights for the physiology of entertainment. Higher average and maximum HR, steeper blood volume signals and quicker changes in HR appear to correlate with higher levels of reported entertainment in children of the age group examined.

Physiological signal features may provide a means for distinguishing between entertaining games and non-entertaining games as well as between gaming activities and game-like physical activities (stomping). This paper examined the hypothesis that physical activity in games reported as entertaining has a quantitatively dissimilar impact on physiology to a non-entertaining game-like form of physical activity. A suitable experiment for controlling and isolating the elements of physical activity from HR, BVP and SC signals was designed. The hypothesis under investigation here is whether there is some kind of physical activity that an entertaining game elicits and a non-entertaining game does not. The control experiment is based on a comparative "fun" analysis between variants of the Bug-Smasher game and the Stomping game which is objectively reported by children as less entertaining than the Bug-Smasher game.

The following protocol was devised to mitigate the difficulties reported in the literature [20] of obtaining accurate data of BVP during physical activity. The HR of children was recorded using a wireless ElectroCardioGram (ECG) device (POLAR s610i) consisting of pulse sensors placed on the chest of the child. The RR intervals (time duration between two consecutive R waves) recorded are automatically converted into HR with an accuracy of ±1 heart beat per minute (bpm). Heart rates were calculated and stored every 5 seconds. The spectral threshold of the Fourier transformation of the BVP was adjusted by observation (where needed) to match those HR recordings. This way we manage to derive RR time intervals out of the BVP signals based on the high accuracy (±1) HR recordings of the ECG device.

Three alternative preference learning mechanisms (LMA, meta-LMA and neuro-evolution) were investigated, demonstrating comparative advantages of evolving ANNs over linear approaches for constructing entertainment models of physical play. Moreover, two alternative signal feature selection (nBest, SFS) methods were applied, demonstrating the benefit of SFS over nBest for obtaining more accurate entertainment models. Specifically, signal feature selection on HR, BVP and SC signal data derived from the main experiment extracted a feature subset including the energy of the high frequency band of HRV (HF), the average HR ($E\{h\}$) and the standard deviation of the inter-beat intervals ($\sigma\{RR\}$). These inputs feed ANNs which correctly predict the reported entertainment preferences of children with a cross-validation accuracy of 79.76% on unseen validation data from the main experiment. Moreover, a performance of 69.64% is obtained when evaluating the ANN trained on data from the main experiment using unseen data from the control experiment. Even though the obtained performance appears low, one has to consider the difficulty of classifying accurately emotions through physiological state [20] and the binomial-distributed probability of this performance to occur at random (**<u>0.0022</u>**). These provide evidence that the evolved ANNs are successful predictors of children's reported entertainment grounded on their physiological state and validate the hypothesis that there are physiological signal features corresponding to physical activity ($E\{h\}$ and $\sigma\{RR\}$) and mental load (HF) that can capture entertainment in physical games.

The proposed entertainment model, in its current state, demonstrates strong evidence that when children are having "fun" during physical play they are engaged more, which is reflected in either increased physical activity or increased mental/emotional load independently of physical effort. On the other hand, there is less evidence about the inverse relation between affect and physical activity in this study since the present model may not be able to recognize positive or negative valence of the affect — e.g. pleasurable excitement from anger. For instance, children irritated by playing Playware games in general may show signs of arousal without increased physical activity, while children motivated to play harder by anger with competitors will exhibit arousal and physical activity but may not be having "fun". To the extent that their physical activity is typical of the Playware game, the ANN model presented here may nevertheless assert that they are enjoying themselves. This may happen when irritated children demonstrate high HF-HRV combined with low $\sigma\{RR\}$ values. Further testing of whether negative valence corresponds to a combination of low mental or stress load and inter-beat homogeneity could refute this possibility inherent in the present model.

As mentioned earlier, the 2-AFC protocol employed in the reported experiments induces "noise" in those cases where the children do not have a strong preference to express. Although the order-effect statistical analysis implies that this noise introduces no bias, the presence of the noise reduces the effectiveness of the machine learning techniques used to construct the predictive models. In view of this, a 4-alternative forced choice (4-AFC) approach could profitably be adopted for future protocol design. Children will choose among the following alternatives: one game is more "fun" than the other (2-AFC), both games are equally "fun", neither game was "fun". This protocol provides the same preference information as 2-AFC for the machine learning process while also making explicit the "no preference" cases concealed by 2-AFC.

A further improvement of the model construction procedure is to use a complete feature selection strategy for the exploration of the feature subset space. SFS with deletion, Sequential Floating Forward Search and Fisher Projection [20] would be suitable candidates to test. On the other hand, a preliminary study in which an exhaustive search of all possible subsets was attempted suggested that the optimal model's performance is only a few percentage points better than the models reported here: in which case the computational effort of a complete search strategy is probably not justifiable.

We believe that the approach to entertainment modeling based on physiological data presented here is general over the majority of action games realizable with Playware. Previous studies have shown that models grounded solely on HR signal features generalize over dissimilar games that collectively cover a large proportion of the features met in Playware action games [5]. Moreover, it is our belief that the entertainment models proposed here may very well be applied to other interactive entertainment systems that include physical activity. Note, however, each game design comprises idiosyncratic entertainment features which may have a significant impact on the child's enjoyment and its physiological consequences. More games, therefore, need to be tested to confirm the generality of the approach.

Individual differences in children's physiology, preferences and playing behavior cause complications in generalizing over subjects and limit the predictive ability of models like those constructed here. This is a fundamental limitation of attempting to construct a model based on combined data from multiple subjects: a game that to one child is exciting and fun may to another be too fast, or too slow. The results presented show an encouraging degree of generalization across individuals to be possible, in that the evolved ANNs do predict children's preferences with reasonable performance. If it were possible to cluster individual players into classes depending on observed playing style, where each class could then have its own model, an easier machine learning problem would result with potentially better predictive performance. On the other hand, capturing sufficient data to train one model per class could pose hard logistical problems.

The predictive models can be used to adapt the game's entertainment features (challenge, curiosity) depending on the variation of the player's individual play and physiological features in real-time (at least — on-line during play) in physical games. The key to this is the observation that the models (e.g. ANNs) relate features to an entertainment value. It is therefore in principle possible to determine what changes to game features (given embedding of the features in the model) will cause an increase in the entertainment value of the game, and to adjust game parameters to make those changes. For further discussion on this future direction the reader may refer to [5].

Acknowledgments

The authors thank Henrik Jørgensen and the children of Henriette Hørlücks School, Odense, Denmark, who participated in the experiments.

The tiles were designed by C. Isaksen from Isaksen Design and parts of their hardware and software implementation were collectively done by A. Derakhshan, F. Hammer, T. Klitbo and J. Nielsen. KOMPAN, Mads Clausen Institute, and Danfoss Universe also participated in the development of the tiles.

This work was supported in part by the Danish Research Agency, Ministry of Science, Technology and Innovation (project no: 274-05-0511).

References

[1] J. Read, S. MacFarlane, C. Cassey, Endurability, engagement and expectations, in: Proceedings of International Conference for Interaction Design and Children, 2002.

[2] R. L. Mandryk, K. M. Inkpen, T. W. Calvert, Using Psychophysiological Techniques to Measure User Experience with Entertainment Technologies, Behaviour and Information Technology (Special Issue on User Experience) 25 (2) (2006) 141–158.

[3] O. Zuckerman, N. Parés, S. Benford, H. H. Lund, Designing interactive environments for outdoors gaming and play, in: Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems: SIGs, Vol. 2, 2006.

[4] H. H. Lund, T. Klitbo, C. Jessen, Playware technology for physically activating play, Artifical Life and Robotics Journal 9 (4) (2005) 165–174.

[5] G. N. Yannakakis, J. Hallam, H. H. Lund, Entertainment Capture through Heart Rate Activity in Physical Interactive Playgrounds, User Modeling and User-Adapted Interaction, Special Issue: Affective Modeling and Adaptation 18 (1-2) (2008) 207–243.

[6] C.-N. Fiechter, S. Rogers, Learning subjective functions with large margins, in: ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000, pp. 287–294.

[7] G. N. Yannakakis, J. Hallam, Towards Capturing and Enhancing Entertainment in Computer Games, in: Proceedings of the 4^{th} Hellenic Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence, Vol. 3955, Springer-Verlag, Heraklion, Greece, 2006, pp. 432–442.

[8] G. N. Yannakakis, H. H. Lund, J. Hallam, Modeling Children's Entertainment in the Playware Playground, in: Proceedings of the IEEE Symposium on Computational Intelligence and Games, IEEE, Reno, USA, 2006, pp. 134–141.

[9] G. N. Yannakakis, J. Hallam, Preliminary Studies for Capturing Entertainment through Physiology in Physical Play, Technical Report TR-2007-5, Maersk Institue, University of Southern Denmark (2007).

[10] G. N. Yannakakis, J. Hallam, H. H. Lund, Capturing Entertainment through Heart-rate Dynamics in the Playware Playground, in: Proceedings of the 5^{th} International Conference on Entertainment Computing, Lecture Notes in Computer Science, Vol. 4161, Springer-Verlag, Cambridge, UK, 2006, pp. 314–317.

[11] D. W. Rowe, J. Sibert, D. Irwin, Heart Rate Variability: Indicator of User State as an aid to Human-Computer Interaction, in: Proceedings of Conference on Human Factors in Computing Systems, 1998, pp. 480–487.

[12] J. J. Goldberger, S. Challapalli, R. Tung, M. A. Parker, A. H. Kadish, Relationship of heart rate variability to parasympathetic effect, Circulation 103 (2001) 1977–1983.

[13] C. Hutt, Exploration and play, in Play and learning, Gardner Press: New York, 1979, pp. 175–194.

[14] M. Hughes, C. Hutt, Heart-rate correlates of childhood activities: play, exploration, problem-solving and day-dreaming, American Psychologist 8 (4) (1979) 253–263.

[15] R. L. Mandryk, M. S. Atkins, A Fuzzy Physiological Approach for Continuously Modeling Emotion During Interaction with Play Environments, International Journal of Human-Computer Studies 65 (2007) 329–347.

[16] N. Ravaja, T. Saari, M. Turpeinen, J. Laarni, M. Salminen, M. Kivikangas, Spatial Presence and Emotions during Video Game Playing: Does It Matter with Whom You Play?, Presence Teleoperators & Virtual Environments 15 (4) (2006) 381–392, in press.

[17] R. L. Hazlett, Measuring emotional valence during interactive experiences: boys at video game play, in: CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, ACM Press, New York, NY, USA, 2006, pp. 1023–1026.

[18] P. Rani, N. Sarkar, C. Liu, Maintaining optimal challenge in computer games through real-time physiological feedback, in: Proceedings of the 11^{th} International Conference on Human Computer Interaction, 2005.

[19] S. McQuiggan, S. Lee, J. Lester, Predicting User Physiological Response for Interactive Environments: An Inductive Approach, in: Proceedings of the 2^{nd} Artificial Intelligence for Interactive Digital Entertainment Conference, 2006, pp. 60–65.

[20] R. W. Picard, E. Vyzas, J. Healey, Toward Machine Emotional Intelligence: Analysis of Affective Physiological State, IEEE Trans. Pattern Anal. Mach. Intell. 23 (10) (2001) 1175–1191.

[21] G. N. Yannakakis, J. Hallam, Entertainment Modeling in Physical Play through Physiology beyond Heart-Rate, in: A. Paiva, R. Prada, R. Picard (Eds.), Proceedings of the International Conference on Affective Computing and Intelligent Interaction, LNCS 4738, Springer-Verlag, Lisbon, Portugal, 2007, pp. 256–267.

[22] M. Liljedahl, S. Lindberg, DigiWall - an audio mostly game, in: International Conference on Auditory Display, London, 2006.

[23] Mixed Reality Lab, Singapore, http://www.mixedreality.nus.edu.sg/.

[24] G. Metaxas, B. Metin, J. Schneider, G. Shapiro, W. Zhou, P. Markopoulos, SCORPIODROME: An Exploration in Mixed Reality Social Gaming for Children, in: Proceedings of ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, 2005.

[25] C. Magerkurth, R. Stenzel, T. Prante, STARS – a ubiquitous computing platform for computer augmented tabletop games, in: Extended Abstracts of UbiComp '03, Springer, 2003, pp. 267–268.

[26] H. Ishii, C. Wisneski, J. Orbanes, B. Chun, J. Paradiso, PingPongPlus: design of an athletic-tangible interface for computer-supported cooperative play, in: CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press, New York, NY, USA, 1999, pp. 394–401.

[27] G. N. Yannakakis, J. Hallam, Game and Player Feature Selection for Entertainment Capture, in: Proceedings of the IEEE Symposium on Computational Intelligence and Games, IEEE, Hawaii, USA, 2007, pp. 244–251.

[28] G. N. Yannakakis, J. Hallam, Modeling and augmenting game entertainment through challenge and curiosity, International Journal on Artificial Intelligence Tools 16 (6) (2007) 981–999.

[29] T. W. Malone, What makes computer games fun?, Byte 6 (1981) 258–277.

[30] G. N. Yannakakis, J. Hallam, Feature Selection for Capturing the Experience of Fun, in: G. N. Yannakakis, J. Hallam (Eds.), Proceedings of the AIIDE'07 Workshop on Optimizing Player Satisfaction, AAAI Press Technical Report WS-07-01, AAAI Press, Stanford, US, 2007, pp. 37–42.

[31] R. Rosenthal, Covert communication in laboratories, classrooms, and the truly real world, Current Directions in Psychological Science 12 (5) (2003) 151–154.

[32] S. M. Pincus, Approximate entropy as a measure of system complexity, Proc. Natl. Acad. Sci. 88 (1991) 2297–2301.

[33] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) 215–220.

[34] J. Doyle, Prospects for preferences, Computational Intelligence 20 (2) (2004) 111–136.

[35] W. Chu, Z. Ghahramani, Preference learning with gaussian processes, in: Proceedings of the 22$^{\text{nd}}$ International Conference on Machine Learning (ICML-2005), 2005.

[36] C. H. Papadimitriou, K. Steiglitz, Combinatorial Optimization: Algorithms and Complexity, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[37] P. Devijver, J. Kittler, Pattern Recognition - A Statistical Approach, Prentice-Hall, Engelwood cliffs, NJ, 1982.

[38] E. Haapalainen, P. Laurinen, P. Junno, H. Tuovinen, J. Roening, Methods for classifying spot welding process: A comparative study of performance, in: IEA/AIE, 2005, pp. 412–421.