Georgios N. Yannakakis
John Hallam

# Preliminary Studies for Capturing Entertainment through Physiology in Physical Games


The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
http://www.mip.sdu.dk
Technical Reports 2007, No 5
June 2007
ISSN No. 1601-4219

# Preliminary Studies for Capturing Entertainment through Physiology in Physical Play

Georgios N. Yannakakis, John Hallam, and Henrik Hautop Lund

Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Campusvej 55, Odense M, DK-5230
{georgios,john}@mmmi.sdu.dk

**Abstract.** This report presents preliminary physical control experiments for capturing and modeling the affective state of entertainment — that is, whether people are having "fun" — of users of the innovative Playware playground, an interactive physical playground. The goal is to construct, using representative statistics computed from children's physiological hear rate (HR) signals, an estimator of the degree to which games provided by the playground engage the players. For this purpose children's HR signals, and their expressed preferences of how much "fun" particular game variants are, are obtained from experiments using games implemented on the Playware playground. Neuro-evolution techniques combined with feature set selection methods permit the construction of user models that predict reported entertainment preferences given HR features. These models are expressed as artificial neural networks and are demonstrated and evaluated on two Playware games and the preliminary control task requiring physical activity. Results demonstrate that the proposed preliminary control experiment is not an appropriate control for physical activity effects since it may generate HR dynamics rather easy to separate from game-play HR dynamics, and allows one to distinguish entertaining game-play from exercise purely on the artificial basis of the kind of physical activity taking place. Conclusions derived from this study constitute the basis for the design of more appropriate control experiments and user models in future studies.

## Keywords

Entertainment modeling, biosignals, intelligent interactive playgrounds, mixed-reality games, artificial neural networks.

## 1 Introduction

Motivated by the lack of quantitative affective models of entertainment in physical play, preliminary experiments for estimating expressed player satisfaction in

real-time through physiological signals measured during gameplay is presented in this paper. Our principal goal in the reported work is to construct a user model of the player of a game — in this case a child playing a physical interactive game — that can predict the answers to which variants of the 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. In this work the model is constructed using machine learning techniques applied to statistical features derived from physiological signals measured during play. The output of the constructed model is a real number in the range $[0, 1]$ such that more enjoyable games receive higher numerical output. This basic approach, defined as *entertainment modeling*, is applicable to a variety of games, both computer [2] and physical [3], using features derived from physiological data or from the interaction of player and opponent measured through game parameters.

Even though entertainment is a highly complicated mental state it is correlated with sympathetic arousal [4, 5] which can be captured through specific physiological signals such as heart rate and skin conductivity, as reported by researchers in the psychophysiological research field [6, 7]. In this paper, we investigate the impact of entertainment on heart rate (HR) signals and attempt to capture HR signal features that correlate with children's expressed entertainment preferences. HR signal data and children's reported preferences between variants of Playware games are obtained through gaming experiments using the Playware playground. HR dynamics are represented by calculating several statistics and regression model parameters from gameplay experimental data, to serve as features for the construction of a user model as described above.

In the main survey experiment presented in [8] 56 children participants were split into two groups of 28 children and each group was assigned to play either Bug-Smasher or Space-Invaders Playware games. By experimental design, each subject played for 90 seconds each against two of the selected game variants, A and B, — differing in the levels of one or more quantitative entertainment factors of challenge, curiosity and fantasy [9] — of either Bug-Smasher or Space-Invaders in all permutations of pairs. Each time a pair of games ('game pair') was finished, the child was asked whether the first game was more "fun" than the second game. The HR of children was recorded in real-time using a wireless ElectroCardioGram (ECG) device consisting of pulse sensors placed on the chest of the child and a number of representative features were extracted from the recorded signals. For details on the Playware platform, the test-beds used (Bug-Smasher and Space-Invaders) and the experiment protocol the reader is advised to refer to [8] since understanding of the following text is strongly dependent on that study.

This report presents a preliminary physical activity control experiment on the investigation between physiology and reported entertainment, being the initial study of the survey experiments reported in [8]. This experiment is focused on the distinction between HR signals corresponding to reported entertainment

preferences in a gaming activity (entertaining or not) and HR signals corresponding to pure (non-game, and non-entertaining) physical activity. Obtained results show that HR dynamics can be used to construct models (of the kind described above) that discriminate well between entertaining game activities and physical exercise. However, the question of whether there is anything in the type of physical activity that is characteristic of an entertaining game remains since the particular control physical activity employed (running in circles), while not entertaining, is also quite different in character to the more variable exertion of a typical Playware game.

## 2 Physical Activity Control Experiment: Entertainment and Exercise

Given the HR signals recorded in the experiments presented in [8] and here the following 14 features are extracted: The average HR $E\{h\}$, the standard deviation of HR $\sigma\{h\}$, the maximum HR $\max\{h\}$, the minimum HR $\min\{h\}$, the difference between maximum and minimum HR $D = \max\{h\} - \min\{h\}$, the correlation coefficient $R$ between HR recordings and the time $t$ in which data were recorded, the autocorrelation (lag equals 1) of the signal $\rho_1$ and the approximate entropy ($ApEn$) [10] of the signal which quantifies the unpredictability of fluctuations in the HR time series (see [8] for further details on $ApEn$). In addition, 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.

Statistical analysis presented in [8, 11] has shown that average HR ($E\{h\}$) is the HR only signal feature that correlates with reported entertainment. Indeed, this feature also corresponds to physical activity and therefore the main conclusion to be drawn is that the more engaging the gameplay, the higher the physical activity (through the aforementioned HR features) and the higher the perceived entertainment for a child.

This raises the question of whether the statistical effects observed genuinely reflect *entertainment* value or merely the tendency of more engaging games to elicit more physical activity. That is, is there anything in the type of physical activity that is characteristic of an enjoyable game or is the analysis just comparing amount of physical activity?

Herein experiments comparing HR signals derived from the Playware physical games with those from pure (non-game) physical exercise are presented. In addition, a neuro-evolution model for discriminating between physical activity and game (entertaining or not) HR signals is proposed in this section.

### 2.1 Experimental Data

In order to investigate the interplay between entertainment and physical activity we asked all children that played the two Playware games (see [8]) to participate in an additional experiment: each child was asked to run around a

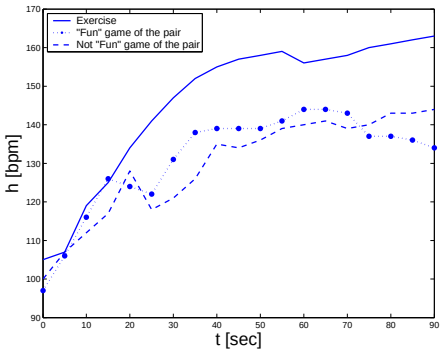
3 m×3 m space for 90 seconds. The assumption here is that this exercise task is a non-entertaining activity for the child. This assumption was supported by most children asking for the time remaining during the task, suggesting a certain level of boredom for the activity. However, children were not asked whether the running task was "fun" or not and it was not compared to any physical game task.

HR signal data is dived in triples each consisting of the game pair played by the child (a game reported as entertaining and a game reported as non-entertaining) and its corresponding running HR signal. Running HR signals obtained cover 34 out of 41 game pairs and 41 out of 48 game pairs played for the Bug-Smasher and the Space-Invaders respectively. In the remaining 14 triples, running HR data was not available partly because four of the children participated in the game experiments refused to just run around in a circle for 90 seconds. This fact strengthens our assumption in that the running task devised should be a non-entertaining form of physical activity. Further loss of running HR signal data was due to hardware failure. This 15% loss of HR signal running data is substantial; however, there is no reason to suppose that both the hardware failure and the unwillingness of specific children to participate have any particular bias with respect to experimental hypothesis. The set of HR time series collected from the 75 correctly recorded triples of tasks (a pair of games and a running task) is used in the analysis presented throughout the remaining of the paper.

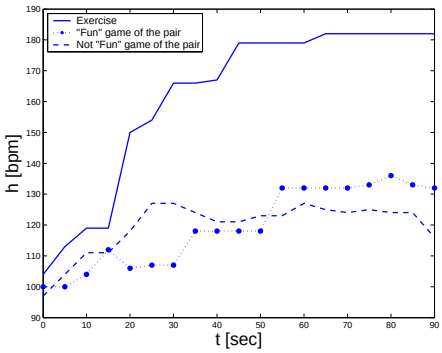
The physical activity control experiment was held two months after the survey experiment described in section 1 (see [8]). It may therefore be affected by variations in the children's physiology mainly due to variations caused by sugar, sleep, hormones and mood [12]. This variation cannot be regarded as insignificant for our purposes and may have an impact on the analysis; however, the focus of this paper is not on the examination of the long-term realistic physiology of subjects (as in [12]) with regards to reported entertainment. The physical activity control experiment described in [8] is designed more carefully to control for physiology's day-dependence since both game and physical activity control tasks are held on the same day.

Some indicative graphs from children's real-time HR recordings during this experiment are presented in Fig. 1. The qualitative features of the exercise HR signal illustrated in Fig. 1 are very similar for all children that participated in this experiment.

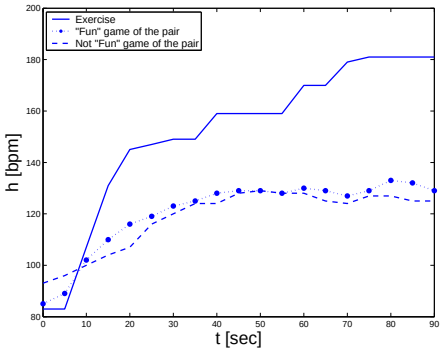
Apparently, children conscientiously followed the rules of the experiment and ran continuously for 90 seconds, generating the HR exercise dynamics presented in Fig. 1. It is clear that constant running exercise generates — in the majority of children — higher $E\{h\}$ values than the corresponding values for Playware gameplay since it requires more demanding physical activity from the child.



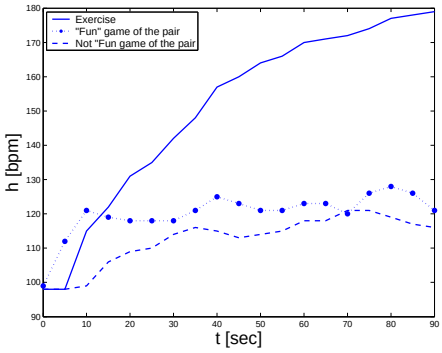

(a) Space Invaders: "Fun" game: *High* challenge, *High* curiosity, not "Fun" game: *High* challenge, *Low* curiosity.




(b) Space Invaders: "Fun" game: *High* challenge, *Low* curiosity, not "Fun" game: *Low* challenge, *Low* curiosity.




(c) Bug-Smasher: "Fun" game: *High* challenge, *Low* curiosity, not "Fun" game: *High* challenge, *High* curiosity.




(d) Bug-Smasher: "Fun" game: *Low* challenge, *High* curiosity, not "Fun" game: *Low* challenge, *Low* curiosity.

**Fig. 1.** Comparative HR signal graphs of four children: exercise (running), entertaining and non-entertaining games

## 3 ANN model construction

The proposed approach to entertainment modeling, introduced in [13] is based on selecting a 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.

As discussed in previous studies [14], preference learning [15] 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 [16] and evolving ANNs. However, 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 on the neuro-evolution mechanism used can be found in [17, 8].

### 3.1 Feature Selection

Two different input feature set selection schemes are used to pick the appropriate feature subset that generates the highest classification performance between preferred and non-preferred games. Given the signal features extracted, the *n Best Features Selection* (nBest) and the *Sequential Forward Selection* (SFS) methods are applied (see [13] for further details). To evaluate the performance of each feature subset considered by each selection algorithm, the available data is randomly divided into training and validation data sets consisting of 2/3 and 1/3 of the data respectively. 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 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 than the performance obtained without it.

## 4 Single Feature Performance

The experiment presented here tests the validation performance of single HR features. Given the selected feature (ANN input) and the available 150 pairs of comparisons (75 pairs of entertaining/non-entertaining game and 75 pairs of entertaining game/exercise task) ANNs are evolved by following the approach presented in [8] and evaluated through the leave-one-out cross-validation method (see section 3.1). The training and validation performance of each of the individual player and game features are presented in Table 1 where features are ranked by validation performance.

**Table 1.** Training and Validation performance (classification accuracy) $P$ and its respective standard deviation ($\sigma\{P\}$) of the 14 features extracted from recorded HR signals. $\{A, B, b\}$ and $\{\beta, \gamma\}$ are parameters of the exponential and quadratic regression models respectively; $s$ is the linear slope of the signal.

| Feature | Validation — $P$ (%) | $\sigma\{P\}$ | Training — $P$ (%) | $\sigma\{P\}$ |
|---|---|---|---|---|
| $ApEn$ | 73.33 | 8.08 | 68.00 | 5.56 |
| $\sigma^2\{h\}$ | 72.00 | 5.29 | 73.00 | 6.08 |
| $E\{h\}$ | 70.00 | 6.92 | 68.66 | 5.13 |
| $\gamma$ | 69.33 | 4.16 | 74.33 | 3.21 |
| $A$ | 68.66 | 8.32 | 72.66 | 2.30 |
| $s$ | 67.33 | 3.05 | 71.33 | 1.52 |
| $max\{h\}$ | 66.76 | 3.05 | 70.33 | 2.08 |
| $D$ | 64.66 | 4.61 | 71.33 | 2.51 |
| $\rho_1$ | 63.33 | 3.05 | 69.33 | 3.05 |
| $R$ | 63.33 | 2.30 | 67.00 | 3.60 |
| $b$ | 60.00 | 8.71 | 64.33 | 3.21 |
| $min\{h\}$ | 59.33 | 6.11 | 65.66 | 2.08 |
| $\beta$ | 58.66 | 12.85 | 65.66 | 10.06 |
| $B$ | 54.66 | 10.06 | 71.00 | 5.29 |

The impact of the approximate entropy ($ApEn$) in reported entertainment is demonstrated in Table 1 since it generates the highest cross-validation performance. Even though the validation performance of $ApEn$ and $\sigma^2\{h\}$ (among others) are statistically equal we rank features according to their average cross-validation classification accuracy. Results obtained show the incapability of a single HR signal feature to successfully predict reported entertainment in Playware games. Given that the best performed feature ($ApEn$) yields a cross-validation performance of 73.33%, it becomes apparent that more statistical features are required to effectively model children's notion of entertainment.

### 4.1 More Features: Selection Method Comparison

This section presents experiments for finding the minimal feature subset that yields the highest classification accuracy in matching the ANNs output with children's reported answers on entertainment in unknown data (validation data set). For this purpose, the nBest and the SFS selection methods are applied and compared. The initial subset (ANN input) for both methods includes the feature that performs best in the single feature experiment: $ApEn$. ANNs are evolved by following the approach presented in [8]. The data is partitioned in training (2/3 of total data) and validation (1/3 of total data) portions and the leave-one-out cross-validation technique is used to obtain the classification performance of the ANNs.

Table 4.1 presents the above-mentioned comparison between nBest and SFS. SFS appears to generate feature subsets that yield higher validation performance than feature subsets generated by nBest. The best cross-validation performance (80.66%; average of 88%, 78% and 76%) is achieved when the ANN input contains $ApEn$ and $E\{h\}$. More HR signal features added in the feature subset do not yield significantly higher classification accuracy (see bottom row of Table 4.1).

The obtained classification accuracy demonstrates the existence of an ANN model that successfully predicts the children's reported entertainment preferences given a child's individual HR signal features: $ApEn$ and $E\{h\}$. However, difficulties in obtaining higher classification accuracy are found in experimental noise in both the recorded features and the children's answers on self reports. Even though comparative "fun" analysis is a reliable and established method for capturing reported entertainment in computer [18] and mixed-reality [3] games, it generates a certain amount of uncertainty in subjects' reported answers. Uncertainty appears when the two games played are not significantly different with regards to the entertainment value they generate for the player and therefore cannot be distinguished. In this circumstance, players appear to express a random preference. This 'dilutes' the data in which genuine preferences are expressed from the point of view of the machine learning algorithm.

**Table 2.** Classification accuracy (%) of random network, nBest and SFS feature selection methods. Random network performance is the average performance of ten random weight value initializations of the network. The random network's input vector consists of the best feature subset that generates the highest cross-validation performance (i.e. $\{R,E\{h\}\}$).

| Random network | nBest | | SFS | |
|---|---|---|---|---|
| $P$ | Feature subset | $P$ | Feature subset | $P$ |
| 41.53 | $\{ApEn\}$ | 73.33 | $\{ApEn\}$ | 73.33 |
| | $\{ApEn,\sigma^2\{h\}\}$ | 68.00 | $\{ApEn,E\{h\}\}$ | **80.66** |
| | $\{ApEn,\sigma^2\{h\},E\{h\}\}$ | 73.33 | $\{ApEn,E\{h\},\gamma\}$ | 80.66 |

### 4.2 Evolved ANN: $\{ApEn, E\{h\}\}$ Feature Subset

This section provides a further analysis of the best classifier presented in Table 4.1. Given the $\{ApEn, E\{h\}\}$ feature subset as inputs, the evolved ANNs were able to correctly match 81.00 % (average of the three training trials; $\sigma = 2.64\%$ obtained through bootstrapping) of children's entertainment preferences while achieving a classification performance of 80.66% ($\sigma = 6.42\%$) on the unseen validation data. Note that the ANN classifies games reported as non-entertaining as a more entertaining activity than the exercise (non-game) tasks with an accuracy of 94.00% and 93.33% on training and validation data portions respectively.

The relation between $ApEn$, $E\{h\}$ and the game's predicted entertainment value ($y$) given by the highest performing ANN found is illustrated in Fig. 2. Note that the three fittest ANNs generated, each trained on different portions of 2/3 of total data, exhibit the same qualitative features of the surface illustrated in Fig. 2.

According to Fig. 2, a physical activity is not entertaining when high $E\{h\}$ values ($E\{h\} > 0.7$) are combined with lower than average $ApEn$ values ($ApEn < 0.5$). Given the experiments presented in this paper, this is a common situation in pure exercise physical activities which are not considered entertaining by children according to our assumption. Highly entertaining games are the ones that correspond to a combination of very high $E\{h\}$ and $ApEn$ values ($E\{h\} > 0.6$, $ApEn > 0.7$).

The high cross-validation accuracy (80.66%) achieved by the evolved ANN indicates that a non-linear formulation of $E\{h\}$ and $ApEn$ can model reported entertainment preferences well. Note that while the t-test for means of paired samples of the $E\{h\}$ values between the games chosen by the subjects as entertaining and the games chosen as non-entertaining shows a significant difference ($t = 2.71$, $P = \underline{\mathbf{0.0041}}$) the respective t-test for the $ApEn$ values shows that the examined HR signal feature values are not significantly different ($t = 0.33$, $P = 0.3705$). Thus the evolved ANN appears to be a reliable predictor for reported entertainment preference within the two Playware games considered and the non-linear combination of input features an essential condition for its ability to do that.

## 5 Conclusions

HR signals obtained show that the running task appears to involve much more physical effort (high $E\{h\}$ values) than the physical effort required in a Playware game, and further that the physical effort involved is different in kind (low $ApEn$ values; high regularity of the HR signal). It follows that the physical activity control experiment presented here may generate HR dynamics rather easy to separate from game-play HR dynamics, and allows one to distinguish entertaining game-play from exercise purely on the artificial basis of the kind of physical activity taking place. It is therefore, in retrospect, not a good control for physical activity effects.

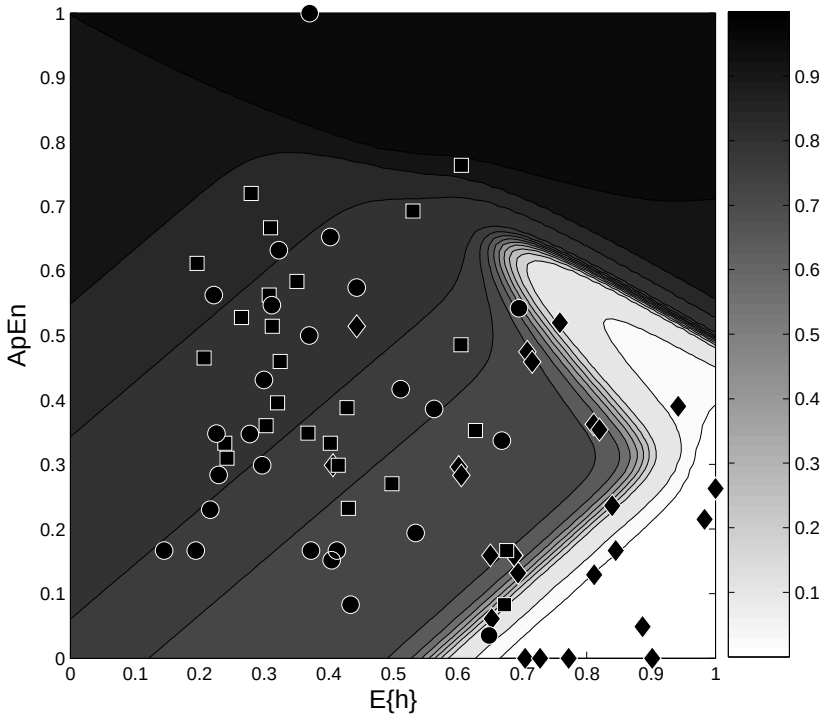



**Fig. 2.** Evolved ANN that yields the best classification accuracy on unknown data (88.00%): ANN output $y$ (entertainment; the darker the higher) with regards to $E\{h\}$ and $ApEn$. Points plotted correspond to the 75 data of the validation set including 25 entertaining games (squares), 25 non-entertaining games (circles) and 25 exercise trials (diamonds).

Consider, though, that one cannot control completely for physical activity effects, since the games being played are *physical* games whose differing enjoyability may naturally be expected to result in different degrees of physical engagement by the player. A better control, therefore, would provide game-like — but non-entertaining — physical activity. A more appropriate experiment for controlling and isolating the elements of physical activity from an HR signal so that features of HR signal corresponding to entertainment become more apparent is presented in [8].

## 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. Read, J., MacFarlane, S., Cassey, C.: Endurability, engagement and expectations. In: Proceedings of International Conference for Interaction Design and Children. (2002)
2. Yannakakis, G.N., Hallam, J.: Towards Capturing and Enhancing Entertainment in Computer Games. In: Proceedings of the $4^{th}$ Hellenic Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence. Volume 3955., Heraklion, Greece, Springer-Verlag (2006) 432–442
3. Yannakakis, G.N., Lund, H.H., Hallam, J.: Modeling Children's Entertainment in the Playware Playground. In: Proceedings of the IEEE Symposium on Computational Intelligence and Games, Reno, USA, IEEE (2006) 134–141
4. Mandryk, R.L., Inkpen, K.M., Calvert, T.W.: Using Psychophysiological Techniques to Measure User Experience with Entertainment Technologies. Behaviour and Information Technology (Special Issue on User Experience) **25** (2006) 141–158
5. Mandryk, R.L., Atkins, M.S., Inkpen, K.M.: A Continuous and Objective Evaluation of Emotional Experience with Interactive Play Environments. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI 2006). (2006) 1027–1036
6. Critchley, H.D., Rotshtein, P., Nagal, Y., O'Doherty, J., Mathias, C.J., Dolan, R.J.: Activity in the human brain predicting differential heart rate responses to emotional facial expressions. Neuroimage **24** (2005) 751–762
7. Zuckerman, M.: Sensation Seeking in Entertainment. In: Phychology of Entertainment. Lawrence Erlbaum Associates Publishers (2006) 367–387
8. Yannakakis, G.N., Hallam, J., Lund, H.H.: Entertainment Capture through Heart Rate Activity in Physical Interactive Playgrounds. User Modeling and User-Adapted Interaction, Special Issue: User Modeling and Affective Computing (2007) (to appear).

9. Malone, T.W.: What makes computer games fun? Byte **6** (1981) 258–277
10. Pincus, S.M.: Approximate entropy as a measure of system complexity. Proc. Natl. Acad. Sci. **88** (1991) 2297–2301
11. Yannakakis, G.N., Hallam, J., Lund, H.H.: 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. Volume 4161., Cambridge, UK, Springer-Verlag (2006) 314–317
12. Picard, R.W., Vyzas, E., Healey, J.: Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Trans. Pattern Anal. Mach. Intell. **23** (2001) 1175–1191
13. Yannakakis, G.N., Hallam, J.: Game and Player Feature Selection for Entertainment Capture. In: Proceedings of the IEEE Symposium on Computational Intelligence and Games, Hawaii, USA, IEEE (2007) 244–251
14. Yannakakis, G.N., Hallam, J.: Feature Selection for Capturing the Experience of Fun. In: Proceedings of the AIIDE'07 Workshop on Optimizing Player Satisfaction. (2007)
15. Doyle, J.: Prospects for preferences. Computational Intelligence **20** (2004) 111–136
16. Fiechter, C.N., Rogers, S.: Learning subjective functions with large margins. In: ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2000) 287–294
17. Yannakakis, G.N., Hallam, J.: Preliminary Studies for Capturing Entertainment through Physiology in Physical Play. Technical Report TR-2007-5, Maersk Institue, University of Southern Denmark (2007)
18. Yannakakis, G.N., Hallam, J.: Towards Optimizing Entertainment in Computer Games. Applied Artificial Intelligence (2007) to appear.