Measuring the Effect of Classification Accuracy on User Experience in a Physiological Game

Gregor Geršak[1], Sean M. McCrea[2] and Domen Novak[2]
[1]*Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia*
[2]*Department of Psychology/Department of Electrical and Computer Engineering, University of Wyoming, Laramie, U.S.A.*

Keywords: Affective Computing, Classification, Computer Games, Physiological Computing, User Experience.

Abstract: Physiological games use classification algorithms to extract information about the player from physiological measurements and adapt game difficulty accordingly. However, little is known about how the classification accuracy affects the overall user experience and how to measure this effect. Following up on a previous study, we artificially predefined classification accuracy in a game of Snake where difficulty increases or decreases after each round. The game was played in a laboratory setting by 110 participants at different classification accuracies. The participants reported their satisfaction with the difficulty adaptation algorithm as well as their in-game fun, with 85 participants using electronic questionnaires and 25 using paper questionnaires. We observed that the classification accuracy must be at least 80% for the physiological game to be accepted by users and that there are notable differences between different methods of measuring the effect of classification accuracy. The results also show that laboratory settings are more effective than online settings, and paper questionnaires exhibit higher correlations between classification accuracy and user experience than electronic questionnaires. Implications for the design and evaluation of physiological games are presented.

1 INTRODUCTION

1.1 Classification in Physiological Games

Computer games represent an important application of physiological computing. In such “physiological games”, physiological measurements are used to detect player boredom, frustration or anxiety and adapt the game difficulty accordingly – decrease difficulty in case of high frustration/anxiety and increase it in case of boredom (Gilleade et al. 2005; Liu et al. 2009; Chanel et al. 2011). This can be done for entertainment purposes, but can also be used in serious applications such as ensuring an appropriate level of exercise intensity in physical rehabilitation (Koenig et al. 2011). Such difficulty adaptation based on physiological measurements has been shown to be more effective than adaptation based on game performance (Liu et al. 2009).

In order to adapt game difficulty, it is necessary to first identify the level of player boredom, frustration or anxiety from the physiological responses. This is generally done with psychophysiological classification algorithms such as discriminant analysis, support vector machines or neural networks (Novak et al. 2012). These algorithms take physiological measurements as the inputs, and output the current state of the user. In most practical cases, this user state has two possible classes (e.g. low / high anxiety) or three possible classes (e.g. low / medium / high frustration). Difficulty is then adapted using simple rules – increase difficulty in case of low frustration, and decrease it in case of high frustration.

No psychophysiological classification algorithm is perfect, and mistakes are always made when trying to identify the state of the user from physiological measurements. The classification accuracy is usually evaluated offline (with previously recorded data) and ranges from as low as 60% (Novak et al. 2011) to as high as 90% (Liu et al. 2009). The “ground truth” for such evaluations is the user’s self-reported level of frustration, anxiety or boredom. This level is obtained via questionnaires that are administered at regular intervals throughout the game. Classification accuracy is then defined as

Geršak, G., McCrea, S. and Novak, D.
Measuring the Effect of Classification Accuracy on User Experience in a Physiological Game.
DOI: 10.5220/0005940300800087
In *Proceedings of the 3rd International Conference on Physiological Computing Systems (PhyCS 2016)*, pages 80-87
ISBN: 978-989-758-197-7

the percentage of times the computer and the user report the same user state. Since this user state is linked to difficulty adaptation via simple rules, classification accuracy can be alternatively defined as the percentage of times the computer and the user agree on how difficulty should be adapted (Novak et al. 2014).

In addition to evaluation with previously recorded data, some studies have also examined classification in real time during gameplay and have mainly achieved accuracies around 80% (Liu et al. 2009; Shirzad and Van der Loos 2016; Liu et al. 2008). These studies reported that users were largely satisfied with difficulty adaptation, suggesting that a classification accuracy around 80% is sufficient for physiological games. However, other questions now arise: Would a lower accuracy have been sufficient as well? Would a higher accuracy have increased user satisfaction further?

1.2 Effect on User Experience

The question of how classification accuracy affects the overall user experience remains surprisingly underexplored in physiological computing. The majority of studies have focused only on a single classification accuracy, as described above. However, for future development of physiological computing, it is critical to know the minimal acceptable classification accuracy for different applications as well as the extent to which classification accuracy improves users' satisfaction with the system. This would allow developers to, for example, determine whether a system is ready for public release or whether the psychophysiological classification accuracy should be improved using additional sensors or better classification algorithms.

A recent study explored the effect of different classification accuracies on user experience in a simulated physiological game (Novak et al. 2014). Study participants played the classic Snake game on the Internet, with no physiological sensors attached. Whenever a new game round began, participants were asked whether they would prefer to increase or decrease difficulty. No option was given to keep difficulty at the same level. The game then simulated a classification accuracy by doing what the user wanted in a certain percentage of situations. For example, to simulate a 100% classification accuracy, the game always changed difficulty as requested by the user. On the other hand, to simulate a 50% classification accuracy, the game changed difficulty in the direction requested by the user in only half of the cases, and changed it in the other direction in the other half of the cases.

The study (Novak et al. 2014) found a significant correlation between classification accuracy and satisfaction with the difficulty adaptation algorithm (r=0.43). However, there was practically no correlation between classification accuracy and in-game fun (r=0.10). This surprising result suggested that the accuracy of user state classification in physiological games does not matter very much.

However, the aforementioned study also suffered from several methodological weaknesses. Most notably, participants were able to drop out in the middle of the study, and dropouts were not included in the analysis. It is thus likely that participants who did not have fun in the game simply dropped out, skewing the results. Furthermore, participants were unevenly distributed into groups, and the questionnaires, which were delivered over the Internet, had not been previously validated.

1.3 Contribution of Our Paper

Due to these methodological weaknesses, we must wonder whether psychophysiological classification accuracy truly has no effect on in-game fun or whether the effect was simply not properly measured. The high dropout rates and uneven group sizes could be easily addressed in a laboratory study. A lab setting supervised by an experimenter would also avoid the potential issue of participants not paying attention to Internet-based instructions (Oppenheimer et al. 2009). Furthermore, it would allow the questionnaires to be delivered either electronically or on paper, which could affect the results.

Our study therefore extends the previous Novak et al. (2014) study to a laboratory setting where dropout rates are very low and participants wear an inactive physiological sensor to simulate an actual physiological computing experience. The same questionnaire items are reused to enable comparison to the previous study. Since little research has been done on measuring the effect of psychophysiological classification, our research contributes to the development of evaluation methods in the field of physiological computing.

2 MATERIALS AND METHODS

2.1 Participants

114 participants were recruited for the study. Due to data collection issues (crashes, incomplete

questionnaires), data from 4 participants were discarded, resulting in 110 valid participants total. They were recruited primarily among students and staff of the University of Ljubljana, Slovenia, with additional participants recruited by word of mouth. There were 74 men and 39 women, mean age 26.4 years, standard deviation 8.5 years.

Participants were divided into two groups: the first 85 participants completed the study with electronic questionnaires and the remaining 25 completed it with paper-and-pencil questionnaires.

Within each group, participants were evenly divided into five subgroups corresponding to five psychophysiological classification accuracies: 33.3% (2/6), 50% (3/6), 66.7% (4/6), 83.3% (5/6) and 100%. Each participant played the physiological game with the classification accuracy assigned to them, as described in the next section.

2.2 Experiment Protocol

The experiment was conducted at the University of Ljubljana, Slovenia. Upon arrival, the experimenter explained to the participant that the goal of the study was to test the performance of a physiology-based game difficulty adaptation algorithm. The game was demonstrated and the participant was allowed to play it briefly. They were told that the game difficulty would be changed according to engagement level measured using a physiological sensor and intelligent machine learning algorithms. They were also told that the game would periodically ask them for their opinion on difficulty adaptation, but that this opinion would only be used to evaluate and further improve the algorithm – it would not affect the decisions taken by the difficulty adaptation.

If the participant consented to participate, a skin conductance sensor was attached to the distal phalanges of the second and third fingers of the nondominant hand. The experimenter pretended to calibrate the sensor to give the impression that the participant's physiological responses would actually be recorded, but in reality the sensor was turned off for the duration of the study. Its purpose was simply to mimic a realistic physiological gaming situation. After the 'calibration', a few pre-game questions were asked (section 2.4).

The participant played the game for seven rounds. At the end of each round except the last one, the participant was asked if they would prefer to increase or decrease difficulty. No option was given to stay at the same difficulty. Once the participant's preference was input, the computer chose whether to change difficulty the way the participant wanted. The probability of doing what the participant wanted was defined by the classification accuracy assigned to the participant (section 2.1). For example, at 66.7% classification accuracy, the computer changed difficulty the way the participant wanted after 4 out of 6 rounds and changed difficulty in the opposite direction after 2 out of 6 rounds.

This computer behaviour was different from what participants were told would happen (that physiological measurements would be used to change difficulty), but the deception was necessary for the purpose of the experiment. Artificially defining the classification accuracy (as the percentage of times the computer does what the player wants) allowed us to study the effects of classification accuracy in a controlled setting. Since the ground truth for psychophysiological classification is generally the player's self-reported state or preference, the artificially defined agreement percentage serves as an appropriate stand-in for actual classification accuracy (Novak et al. 2014).

After the seventh round, the participant completed a questionnaire about their overall game experience (section 2.4). They were then debriefed about the true purpose and protocol of the study.

2.3 The Game

The game was reused from the Novak et al. (2014) study. It is a variant of the classic Snake game where the player controls a snake that moves across the screen at a certain speed (Fig. 1). The player cannot slow down or speed up the snake, but can turn it left or right with the left and right arrows on the keyboard.

The game is divided into rounds. Each round begins with a very short snake and a piece of food at a random position on the screen. When the snake collides with the food, the food disappears, the player gets 100 points, the snake grows longer, and a new piece of food appears at a random position. When the snake collides with itself or the edge of the playing field, it dies and the round ends.

In the first round, the game begins at a random difficulty level between 2 and 6. The difficulty level affects the speed with which the snake moves across the playing field: it needs approximately 7 seconds to cross the entire field at level 1, 2 seconds at level 6, and 1 second at level 10. The difficulty is increased or decreased by one level after each round as described in the previous section.

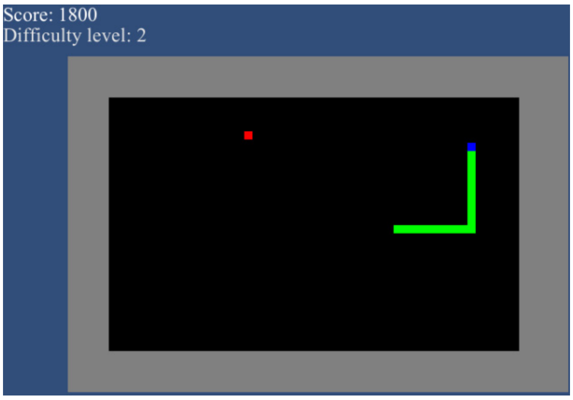

Figure 1: The Snake game. The playing field is colored black and bordered with gray. The snake is green with a blue head. The food (single square) is red.

2.4 Questionnaires

To directly compare our results with those of the previous Novak et al. (2014) study, the same questions were reused even though they have not been extensively validated. Participants were divided into two groups as described in section 2.1: one group used electronic questionnaires and the other used paper-and-pencil questionnaires. In the electronic questionnaires, multiple-choice questions were answered by clicking on the desired answer and visual analog scales were answered by using the mouse to drag a horizontal slider that started in the exact middle of the scale. In the paper questionnaires, multiple-choice questions were answered by circling the desired answer and visual analog scales were answered by marking the desired answer on a horizontal line. For both types of visual analog scales, answers were converted to values between 0 and 100.

Before playing the game, participants were asked about their gender, age, and how often they play computer games (options: "never", "less than 1 hour/week", "1-2 hours/week", "2-5 hours/week", and "more than 5 hours/week"). Participants were also asked how difficult they prefer games to be and how easily frustrated they are. These two questions were answered using a visual analog scale marked "not at all" on one end and "very difficult" or "very easily" on the other end.

After playing the game, participants were asked:

- “How fun was the game?”
- “How frustrating was the game?”
- “How satisfied are you with the decisions of the difficulty adaptation algorithm?”
- “Would you recommend this difficulty adaptation algorithm for practical use?”
- “Would you play this game again with the same difficulty adaptation algorithm?”

The first three questions were answered using visual analog scales and the last two were answered yes or no.

2.5 Data Analysis

Spearman correlations between classification accuracy and answers to post-game questions as well as between pre- and post-game answers were calculated. The analysis was done separately for the group that used electronic questionnaires and the group that used paper-and-pencil questionnaires. The threshold for statistical significance was set at p = 0.05.

3 RESULTS

The correlation between *classification accuracy and satisfaction with the difficulty adaptation* algorithm was significant for both the electronic questionnaire group ($\rho = 0.58$, $p < 0.001$) and for the paper-and-pencil group ($\rho = 0.74$, $p < 0.001$). The correlation between *classification accuracy and in-game fun* was significant for the paper-and-pencil group ($\rho = 0.53$, $p = 0.009$), but not for the electronic group ($\rho = 0.21$, $p = 0.06$). These correlations are illustrated in Figures 2 and 3. The correlation between classification accuracy and in-game frustration was not significant for either group (electronic group: $\rho = -0.19$, $p = 0.08$; paper-and-pencil group: $\rho = 0.24$, $p = 0.24$).

The percentage of players that would recommend a particular difficulty adaptation algorithm for practical use or play the game again with the same difficulty adaptation algorithm is shown in Table 1.

Finally, in the group that used electronic questionnaires, age was negatively correlated with how often the participant plays computer games ($\rho = -0.34$, $p = 0.002$). No other significant correlations between pre- and post-game questionnaire answers were found in either group. In particular, no influence of age or gender was found on user experience with our specific game.

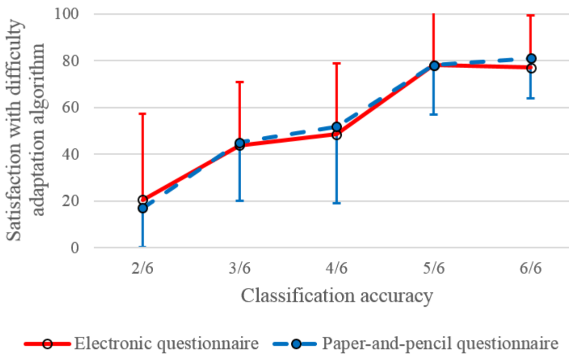

Figure 2: Satisfaction with the difficulty adaptation algorithm, measured using electronic questionnaires (85 participants) and paper-and-pencil questionnaires (25 participants). Error bars represent standard deviations.

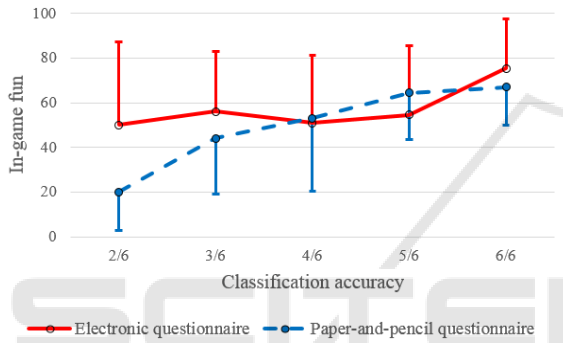

Figure 3: In-game fun as a function of classification accuracy, measured using electronic questionnaires (85 participants) and paper-and-pencil questionnaires (25 participants). Error bars represent standard deviations.

4 DISCUSSION

4.1 Minimum Acceptable Accuracy

Classification accuracy was strongly correlated with satisfaction with the difficulty adaptation algorithm, demonstrating that participants clearly do notice the accuracy of a psychophysiological classification algorithm in a physiological game. Both types of questionnaires showed a strong jump in user satisfaction when classification accuracy increased from 66.7% (4/6) to 83.3% (5/6), as seen in Figure 2. Similarly, Table 1 indicates that both groups have a positive user experience with psychophysiological classification accuracies of 83.3% (5/6) and 100%, but not with lower accuracies.

Based on these results, we can posit that the minimum acceptable accuracy in such a physiological game is approximately 80%, which is similar to the threshold suggested by the previous Novak et al. (2014) study. Lower accuracies result in lower satisfaction with the difficulty adaptation. Designers of physiological games should thus aim to achieve a psychophysiological classification accuracy of at least 80% in order for their game to be accepted by end-users.

Interestingly, the difference between accuracies of 83.3% and 100% is small, as seen in Figures 2-3 and Table 1. This suggests that users are willing to accept occasional mistakes made by the psychophysiological classification algorithm as long as a single mistake is not critical enough to spoil the entire experience. Similar observations have been made in related fields such as games controlled with active brain-computer interfaces (van de Laar et al. 2013). This suggests that efforts to increase psychophysiological classification accuracy above 90% are likely worthwhile only if they require a small investment. For example, adding a simple skin temperature sensor to increase accuracy from 90% to 95% may be worthwhile since the sensor is inexpensive, unobtrusive and requires minimal signal processing. Conversely, adding a multichannel EEG system for the same increase in accuracy would likely not be worth it, as the small increase in user satisfaction with the classification accuracy would be offset by the increased cost, signal processing complexity, and setup time.

Table 1: Percentage of participants who answered "yes" to the two post-game questions, as a function of classification accuracy. Presented separately for participants who used electronic questionnaires (N=85) and participants who used paper-and-pencil questionnaires (N=25).

	Questionnaire type	Classification accuracy				
		2/6	3/6	4/6	5/6	6/6
Would you recommend this difficulty adaptation algorithm for practical use?	Electronic	25%	43.8%	47.3%	95.2%	87.5%
	Paper-and-pencil	0%	40%	40%	80%	80%
Would you play this game again with the same difficulty adaptation algorithm?	Electronic	41.7%	43.8%	68.4%	76.2%	87.5%
	Paper-and-pencil	20%	20%	60%	60%	80%

4.2 Differences between Experiment Settings and Questionnaire Types

4.2.1 Comparison of Results

The effect of psychophysiological classification accuracy on user experience in the Snake game has now been evaluated in three ways: Novak et al. (2014) studied it online with electronic questionnaires while we examined it in a lab setting with electronic and paper-and-pencil questionnaires.

All three approaches found significant correlations between classification accuracy and satisfaction with the difficulty adaptation algorithm. However, the correlation was weakest in the online study of Novak et al. (ρ = 0.43), medium in our electronic questionnaires (ρ = 0.58) and highest in our paper-and-pencil questionnaires (ρ = 0.74). Interestingly, the correlation between classification accuracy and in-game fun was not significant for either the previous Novak et al. study (ρ = 0.10) or our own group that used electronic questionnaires (ρ = 0.21). However, it was highly significant for our paper-and-pencil questionnaire group (ρ = 0.53).

The link between classification accuracy and user experience in our laboratory study was stronger than in the online setting of the previous Novak et al. (2014) study. Additionally, though results are not completely reliable due to the unbalanced number of participants, the paper-and-pencil questionnaires indicated a stronger relationship between classification accuracy and in-game fun than electronic questionnaires. This is somewhat surprising, as the measured relationship between classification accuracy and satisfaction with the difficulty adaptation algorithm is similar for both types of questionnaires (Figure 2).

Nonetheless, assuming that results of paper-and-pencil questionnaires are the more valid ones (as discussed in the next section), they clearly show that increasing the accuracy of psychophysiological classification increases the amount of fun that players have in a physiological game. This is contrary to the surprising result of the previous Novak et al. (2014) study, which found only a minor effect of classification accuracy on in-game fun and thus asked whether increasing classification accuracy is even worthwhile. Our study instead shows that classification accuracy has a strong effect on in-game fun, but that it can be difficult to measure properly. We must thus ask ourselves: what is the reason for the difference between online and lab settings, and between electronic and paper-and-pencil questionnaires?

4.2.2 Possible Explanations

We believe that our lab setting produced better results than the previous online setting of Novak et al. (2014) due to the much lower dropout rate (3.5% in our study, 40% in the previous study). Authors of the previous online study acknowledged that their high dropout rate likely skewed results, as participants who did not enjoy the game simply quit playing rather than fill out the final questionnaires. This would explain the generally better results of our questionnaires, as a more representative sample of user experience has been obtained. A second possible explanation is that participants in the online study may not have paid attention to the instructions, a problem common to all Web-based research (Oppenheimer et al. 2009).

The difference between paper-and-pencil questionnaires and electronic questionnaires in our study is more surprising. Having examined the individual results in detail, we believe that it is due to a weakness in the electronic visual analog scales. Specifically, the slider of the electronic visual analog scale for in-game fun is initially set to the exact middle value, and the participant can adjust the answer by moving the slider. 10 of 80 participants did not move the slider at all; 17 of 80 moved it to the very far left or the very far right. Conversely, in the pencil-and-paper version, only 1 of 25 participants placed the mark at approximately the middle of the scale, and 2 of 25 placed the mark at approximately the far left. We performed a follow-up qualitative examination of the results of the Novak et al. (2014) study and found a similar trend among the 261 participants there: many either left the slider at the default value or dragged it to one extreme. This issue can be at least partially avoided by not providing a starting setting for the electronic visual analog scale and by discouraging participants from selecting extreme values.

Here, we also acknowledge that the observed difference between electronic and paper-and-pencil questionnaires requires deeper experimental investigation. We had originally planned to use only electronic questionnaires, but later added paper-and-pencil questionnaires so that we could check the effect of questionnaire type. However, the two participant groups have very different sizes, and more participants should be tested with paper-and-pencil questionnaires to ensure that the ‘better’ result of such questionnaires is not simply a statistical fluke.

4.2.3 Implications

The differences between questionnaires and settings have important implications for future studies of user experience in physiological computing. First, our study shows that the experimental paradigm of simulating (artificially predefining) classification accuracy does allow user experience to be studied in physiological games, as it captures the effect of psychophysiological classification accuracy on in-game fun. Since no actual physiological sensors are strictly necessary for this, it is tempting to perform studies online and obtain a large number of participants. However, it is critical to avoid high dropout rates, as they may bias results. This could be done using a monetary reward, as is common on, e.g., Amazon Mechanical Turk (Paolacci et al. 2010). On the other hand, such monetary rewards could decrease the ecological validity of the study, as participants should play computer games for fun rather than money.

Furthermore, though the different items on the questionnaire appear to be valid for this physiological game, the method of their presentation is critical – as seen by the difference between electronic and paper-and-pencil questionnaires. Admittedly, the results are not entirely reliable due to the unbalanced number of participants in the two questionnaire groups. Nonetheless, if future studies of user experience in physiological games wish to use similar questionnaires, they should carefully evaluate their method of presentation in different settings. Alternatively, future studies could instead use better-established user experience questionnaires such as the Intrinsic Motivation Inventory (McAuley et al. 1989) or the Game Experience Questionnaire (http://www.gamexplab.nl). Finally, future studies could omit self-report questionnaires entirely and instead use more objective methods of measuring the effect of a physiological game. For instance, in a physiological game for physical rehabilitation (Koenig et al. 2011; Shirzad and Van der Loos 2016), the outcome of rehabilitation could be used as an objective metric of the game's effectiveness and should strongly depend on the game's ability to correctly identify the user's psychological state.

4.3 Other Physiological Games

While our study has been performed only with a single physiological game, we believe that the results would generalize to other games to some degree. They would best generalize to physiological games in which difficulty is always increased or decreased by one level, as in the Snake game. They should also, to some degree, generalize to games that allow difficulty to stay the same or to games that can change difficulty by more than one level. However, the minimum necessary classification accuracy would likely increase in games in which a single erroneously made change can have very negative consequences (e.g. player's avatar immediately dies due to a mistake made by the classification algorithm).

A significant disadvantage of our study is that it only examines user experience in a single brief gameplay session. If users play a game for multiple longer sessions, they may become more aware of the behavior of the physiological game, which could also alter their perception of the classification algorithm (Fairclough 2009). All other previous studies of user experience in physiological games have also only focused on single sessions (Liu et al. 2009; Shirzad and Van der Loos 2016; Liu et al. 2008; Novak et al. 2014), and multisession studies are critical to characterize the evolving relationship between the physiological game and the player.

5 CONCLUSIONS

Our study found a significant effect of psychophysiological classification accuracy on user experience in a physiological game. Both in-game fun and satisfaction with the difficulty adaptation algorithm increased with classification accuracy. The minimum acceptable accuracy for a practical physiological game is approximately 80%, as lower accuracies result in a poor user experience.

We also found that a laboratory setting captures user experience better than an online setting, and that paper-and-pencil questionnaires are more effective than electronic questionnaires when visual analog scales are involved. These results contribute to the development of physiological game evaluation methods, though significant work still needs to be done in order to determine optimal methods of measuring user experience in physiological computing.

ACKNOWLEDGEMENTS

The authors would like to thank all participants as well as the student research assistants who supported data collection.

REFERENCES

Chanel, G. et al., 2011. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans*, 41(6), pp.1052–1063.

Fairclough, S.H., 2009. Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2), pp.133–145.

Gilleade, K., Dix, A. and Allanson, J., 2005. Affective videogames and modes of affective gaming: assist me, challenge me, emote me. In *Proceedings of DiGRA 2005*.

Koenig, A. et al., 2011. Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 19(4), pp.453–64.

van de Laar, B. et al., 2013. How much control is enough? Influence of unreliable input on user experience. *IEEE Transactions on Cybernetics*, 43(6), pp.1584–1592.

Liu, C. et al., 2009. Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. *International Journal of Human-Computer Interaction*, 25(6), pp.506–529.

Liu, C. et al., 2008. Online affect detection and robot behavior adaptation for intervention of children with autism. *IEEE Transactions on Robotics*, 24(4), pp.883–896.

McAuley, E., Duncan, T. and Tammen, V. V., 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: a confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, 60(1), pp.48–58.

Novak, D. et al., 2011. Psychophysiological measurements in a biocooperative feedback loop for upper extremity rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 19(4), pp.400–410.

Novak, D., Mihelj, M. and Munih, M., 2012. A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing. *Interacting with Computers*, 24, pp.154–172.

Novak, D., Nagle, A. and Riener, R., 2014. Linking recognition accuracy and user experience in an affective feedback loop. *IEEE Transactions on Affective Computing*, 5(2), pp.168–172.

Oppenheimer, D.M., Meyvis, T. and Davidenko, N., 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), pp.867–872.

Paolacci, G., Chandler, J. and Ipeirotis, P.G., 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), pp.411–419.

Shirzad, N. and Van der Loos, H.F.M., 2016. Evaluating the user experience of exercising reaching motions with a robot that predicts desired movement difficulty. *Journal of Motor Behavior*, 48(1), pp.31–46.