The Effect of Multiplayer Game Modes on Inter-Player Data for Player Experience Modelling

Alexander Brooke
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, United Kingdom
0009-0009-5907-9044

Matthew Crossley
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, United Kingdom
0000-0001-5965-8147

Huw Lloyd
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, United Kingdom
0000-0001-6537-4036

Stuart Cunningham
School of Computer and Engineering Sciences
University of Chester
Chester, United Kingdom
0000-0002-5348-7700

***Abstract*—Research into social compliance, emotional contagion and behavioural synchronicity shows promise for various avenues of work concerning human-computer interaction, and a wider understanding of emotion. Despite their relevance, few studies have applied findings from these domains to player experience modelling in a multiplayer game, in itself having applications in entertainment, education and healthcare. Further to this, of the little work making use of inter-player data to model aspects of player experience, none considers the differences that may be found across common multiplayer game modes. This work therefore makes use of data collected across players in a series of common multiplayer game modes, considering the utility of inter-player data for predictive modelling using artificial neural networks in each. Results suggest that approaches modelling measures of players' experiences in terms of discrete emotion intensities are best made using their own facial expressions in nearly all circumstances, but past this, facial expression data from team based and competitive game modes shows the greatest promise. Considering the additional data separations available to team-based gameplay, we find that data collected from players on an opposing team shows greater utility for prediction of target player experience than data collected from a player on the same team. Regarding this, we make suggestions for the most applicable avenues for future research into the utilisation of inter-player data for emotional modelling.**

***Index Terms*—multiplayer, emotion, experience modelling**

I. Introduction

The prediction of emotional and experiential response to affective stimulus is of great interest in the field of games research, with human-computer interaction key to the development and implementation of user-responsive games for entertainment, education and healthcare [1]. Despite this, the overlap between research into group emotion recognition and multiplayer games research is largely underserved. Works considering emotion contagion and behavioural synchronicity, for example, explore behavioural, affective, and psychophysiological changes across pairs or groups of individuals [2]–[4], the effects of which are expected to greatly impact multiplayer gaming experiences.

979-8-3315-9919-5/25/$31.00 ©2025

Despite this, games research relating to group emotion often seeks to emulate effects such as emotion contagion for the improvement of tasks such as creating engaging interactions with non-player characters [5], rather than considering such effects for their utility in emotion recognition. Recent work considering the utility of inter-player data for player experience modelling in a multiplayer game has however suggested that the facial expressions of non-target players in a multiplayer setting can indeed be used to improve the accuracy of predictive models of target player affect [6]. Whilst this aligns with the success of approaches utilising group data in the Emotion Recognition in the Wild challenge in 2018 [7], preliminary work by the authors of this study shows less promising results, with improvement using inter-player data seen only for groups that knew each other well, ahead of time. Consideration of the differences between these studies prompts this further research into the utility of inter-player data, with maintained study design, and extended functionality in a test-bed game, used to test the potential for using inter-player data to model target player affect in multiple common game modes in a multiplayer game.

As a part of this work, we consider the effect of three common multiplayer game modes on the utility of inter-player data for emotional intensity estimation, comparing this to the utility of a player's own affective response. We therefore seek to respond to the following research questions:

RQ1 What are the effects of common multiplayer game modes on the predictive power of models utilising players' own affective responses and those of other players in a shared experience to predict their self reported emotional experience?

RQ2 How does affective response data collected from other players in a shared experience compare to that collected from the target player when used for the prediction of self reported emotional experience in common multiplayer game modes?

II. Related Work

Previous work considering the effect of common game modes on the utility of data collected for modelling player experience shows promise, with recent work providing a diverse corpus of gameplay data collected from first person shooter games, across various game modes and games [8]. Preference learning on this dataset provided an indication that deathmatch style gameplay provides data with the greatest utility for the prediction of engagement, over singleplayer and battle royale gameplay. Various works by Kivikingas and Ravaja then consider the difference between cooperative and competitive play, with a greater focus on the utilisation of physiological data over that collected purely from the game system [9], [10]. Across these works, a greater level of social compliance (similarity in physiological signals) was found in pairs of players playing competitively, compared to those playing cooperatively. Findings also suggested that male audiences are more likely to prefer competitive to cooperative gameplay, with a greater level of emotional responses observed during competitive play. This provides a potential explanation for the success seen in previous work considering the utility of inter-player data [6], with data collected from a competitive game and all-male participants potentially providing optimistic results for the generalisability of findings to further genres of game. By contrast, work considering multiple game modes is often confined to data collected from individual players, with exploration into the utility of data specific to the multiplayer scenario overlooked in existing literature.

Experience modelling in previous work considers various measures of player affect (such as arousal and valence) [11], [12] and game related measures (such as competence and engagement) [13], [14], although few concerning discrete emotional labels make use of validated self-reported measures. Recent consideration of the Discrete Emotions Questionnaire (DEQ) [15] for games contexts suggests viability in its use for games user research [16], with work on facial expression annotation utilising the 32 emotional descriptors that the questionnaire assesses [17] in later work.

III. Method

A. Participants

A total of 20 participants were recruited to take part in this study following institutional ethical approval. Participants were recruited via convenience sampling, with participation advertised in university shared areas. Due to this, participants were all young adults (18 aged 18-24, 2 aged 25-34), predominantly male (17 male, 2 female, 1 other), and predominantly white (12 White or White British, 3 Asian or Asian British, 3 Black, African, Caribbean or Black British, and 2 Mixed or Multiple ethnic groups). All bar four participants reported playing three or more hours of video games per week, and the majority reported playing games most commonly on their computer (17 computer, 1 console, 2 mobile phone/tablet).

All eligible applicants (those aged 18 or above, and with the required hardware- Windows computer, microphone and webcam) were provided with a participant information sheet and consent form. Following informed consent, participants were allocated to groups of four randomly (allowing for availability), and took part in the study remotely, each connecting to the rest of their group via an internet connection on their own hardware.

B. Stimuli

Participants were asked to play through a series of levels of custom test bed game "*Colour Rush*". *Colour Rush* has been used in previous work, allowing for multiplayer gameplay, with built-in data collection. The game involves players collecting coins by completing colour mixing tasks, exploring and interacting with the game world, or engaging in combat.

Colour Rush is a 2D semi-top-down game, developed in the Unity game engine [18]. Players control characters that can move around, sprint, crouch, and make use of a lasso tool and flamethrower, all through the use of a keyboard and mouse. Each level is procedurally generated with preselected level generation settings, and features paint blobs that the players can drag around. Paint blobs can be manipulated using environmental tools present in each level, to mix primary colours and split secondary colours. These can then be used to complete "orders" that require paint of specific colours to be dropped off at a designated drop-off zone. Completing orders grants players with the largest number of coins, although these can also be found by exploring the level, and breaking barrels or fighting others using the flamethrower. During each level, player's scores (how many coins they have collected) are presented at the left of the screen, along with a player ranking and group star total at the right of the screen. Group stars are awarded based on the number of coins collected across the entire group. An example section of a level of *Colour Rush* is presented in Figure 1.

Fig. 1. The starting area of a cooperative level of *Colour Rush*. Pictured are: a player, the level's drop off point with a colour task, a paint blob, the main game UI.

Levels of *Colour Rush* last for exactly three minutes, after which participants are provided with an embedded version of the DEQ to report their emotional response. For this, the game prompts participants with the question "Whilst playing the previous level, to what extent did you experience these emotions?", with responses provided for the 32 emotional

descriptors given in the DEQ on a series of seven-point Likert scales from "Not at all" to "An extreme amount".

A stereotypical level from a previous study utilising *Colour Rush* [19] was adapted to form four similar but distinct levels, with variations in level size and the number of coins found in barrels (incentivising and de-incentivising exploration and combat). The game was adapted to provide players with each of these levels three times, covering each of the game mode options new to this version of *Colour Rush*.

These were:

1) **Competitive** Players play the game without teams. They may still work together to move paint blobs around the level, but collect coins independently. The group star ratings at the end of the level and side of the screen are hidden, as players are not incentivised to work as a team in this game mode. The text "Free for All" is displayed at the top of the screen.
2) **Team** Players play the level in two-person teams. These are denoted by players sharing the same player colour, and sharing a coin total. The coin total is shared in real time as opposed to summed at the end between partners to reduce in-team competition during the level. The group star ratings at the end of the level and the side of the screen are again hidden. In this game mode, players on the same team cannot set each other alight. The text "Teams" is displayed at the top of the screen. Teams are consistent across levels, allowing players on the same team to build familiarity.
3) **Cooperative** - Players play the level as a group. Players share the same player colour and coin total. Instead, the group star ratings at the right of the screen and the end of the level are presented as a qualifier of the group's success. The text "Cooperative" is displayed at the top of the screen, as can be seen in Figure 1.

C. Data Collection

After connecting using the game's network functionality, participants were placed into a tutorial level, through which they could learn the controls for the game. This was followed by the 12 described levels of *Colour Rush*, repeating the four generated levels across each of the game modes. Each group played through the levels in a randomised order to reduce order bias. The principal investigator was available to participants in a Microsoft Teams call during the tutorial and following the session, but was not a part of the call during the main portion of the study to reduce observer effect. Participants remained in the call for the entirety of the session, allowing for vocal communication.

In line with previous work in this area, data used in this study describes the facial expressions of participants over the course of each level of gameplay [6], [20]–[22], with previous work detailing the importance of facial expressions in shared emotional effects [23]. Footage was collected at 1280x720p and 10 frames per second, and was processed using OpenFace 2.2 [24], which reported 17 facial action unit (FAU) intensities per frame of footage. A confidence metric provided by OpenFace for each frame was then used to treat the dataset, retaining only frames with a confidence value of 75% and above, inline with previous work [6]. A total of 7.53% of data was removed in this way, with this evenly distributed across the three gamemodes ($n = 3$, $SD = 0.63\%$). Remaining FAU ratings were condensed to a series of intensity ratings for Ekman's six basic emotions (anger, disgust, fear, happiness, sadness and surprise) [25], calculated as the mean average of prototypical FAUs described by Du et al., [26], to improve interpretability as in previous work [22]. From these, a series of feature groups were created summarising the facial expression intensities of each player (treating each as the "target player") and the other players in their group (the "non-target players"), over the 20 players and 12 levels. These were:

1) **Target Player Affect (Self)** The mean and standard deviation of the six basic emotion intensities collected from the target player. This resulted in 240 series of 12 metrics.
2) **Non-Target Player Affect (Group)** The average of the mean and standard deviation metrics collected from all non-target players in the same level. This resulted in 240 series of 12 metrics.
3) **Non-Target Player Affect (Partner)** The mean and standard deviation of the six basic emotion intensities collected from the target player's partner in the teams game mode. This resulted in 80 series of 12 metrics, due to only being applicable in the teams mode.
4) **Non-Target Player Affect (Opposition)** The average of the mean and standard deviation metrics collected from the two players opposing the target player in teams mode levels. This resulted in 80 series of 12 metrics, again due to only being applicable in the teams mode.

D. Target Emotions

The Likert scale ratings for each emotional descriptor in the DEQ collected after each level were used to calculate intensities for each DEQ subscale (anger, anxiety, desire, disgust, fear, happiness, relaxation and sadness), averaging ratings across each subscale's related descriptors, as in the original literature [15]. From this, we consider the distribution of emotional responses provided in response to *Colour Rush*, as seen in Figure 2.

Interestingly, distributions reflect the similar response patterns seen in previous work utilising *Colour Rush* [19], with desire, happiness and relaxation having the greatest variance. We expect this to reflect the gameplay style of *Colour Rush*, although further work may concern the ability of each DEQ subscale to adequately describe gameplay experiences, as it has for recalled gameplay [16]. As in previous work utilising *Colour Rush*, we focus this study on the predictive modelling of the desire, happiness and relaxation subscales from the DEQ, terming these *'target emotions'*, with greater insight for the remaining subscales expected from further genres of game that may more appropriately present a challenge for predictive modelling.

Ahead of further analysis, all data was normalised to a 0-1 scale using min-max normalisation.

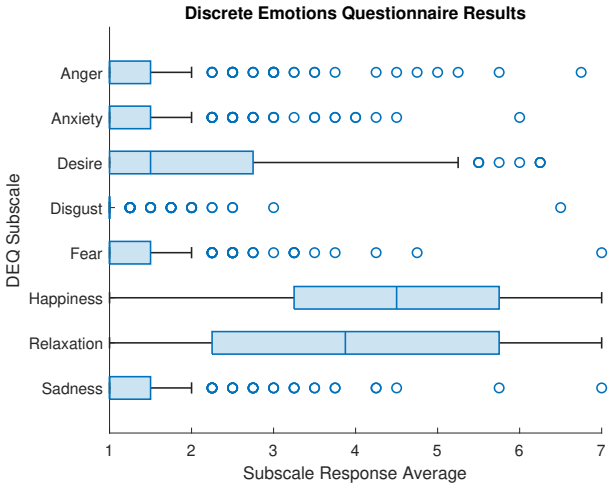

Fig. 2. Ratings for the Discrete Emotions Questionnaire subscales of anger, anxiety, desire, disgust, fear, happiness, relaxation and sadness. Each rating is calculated as the average of four responses related to each subscale, provided after each level of *Colour Rush* by each player.

E. Feature Group Performance

The potential of each feature group for the prediction of the target player's emotional response to each game mode was determined through predictive modelling using artificial neural networks, applying a methodology similar to that of Shaker et al., [20]. Data was split by participant group, making use of leave-one-group-out (LOGO) cross validation ahead of feature selection to increase the generalizability of results and reduce data leakage between training and test data for final performance evaluations. Internally to this, we use sequential forward selection (SFS) to select relevant features from each feature group, as seen in various previous works [11], [27].

The SFS algorithm iteratively builds feature sets by testing the performance of models using all features in a retained feature set and each of the non-retained features, adding the most successful additional feature to the retained set at each iterations until improvements are not gained through the addition of further features. For this, we use the mean squared error (MSE) between emotional intensity predictions and the true values in retained test data as a measure of performance, treating this as continuous data due to the large number of possible values attainable for each DEQ subscale.

Feature subset performance internal to the SFS algorithm also made use of a further LOGO cross validation loop, utilising each of the four groups not retained for testing in the outer fold as a test set. As in previous work, for the first iteration of this algorithm (in which there are no retained features), performance for each individual feature was determined through the use of a single layer perceptron (SLP) trained to predict the target emotional intensity, with further iterations testing each feature subset using a multi-layer perceptron (MLP) with a single two-neuron hidden layer [20], [28]. This resulted in four feature sets per outer fold, per feature group, for each target emotion and game mode.

A more optimal network topology for each feature set was provided through grid search of all neuron counts from 1-30 and 1-10 for each of the two hidden layers in a further series of MLPs, as in previous work [27]. This was conducted internally to the outer cross validation loop, with final performance evaluations for each feature set provided by models making use of the most performant network topology per feature set and trained on all data not retained for testing in the appropriate outer fold. Final evaluations therefore describe the performance of models predicting emotional intensity for completely unseen data.

F. Data Analysis

To appropriately compare model performance (and therefore the potential for use from each feature group) between game modes, we utilise root relative square error (RRSE) as the final measure of performance, using the following standard formula:

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(a_i - p_i)^2}{\sum_{i=1}^{n}(a_i - \bar{a})^2}}$$

Where:

- a denotes the actual values from the test set.
- p denotes the predicted values for the test set.
- n denotes the number of values in the test set.

Use of RRSE scales each MSE statistic using the variance of the test data used for the final performance evaluation of the corresponding model. Through this, models tested on retained data with a lower variance are not favoured more generously, meaning comparison between RRSE values across game modes can be made. Error ratings for each model are therefore standardised, with a value of one representing the error of a mean predictor trained on the output data itself. Improvements seen through lower RRSE values therefore represent models making use of data with a greater potential to be used to estimate DEQ intensity ratings.

With potential differences between each outer fold of data (each group of players), we compare model performances for each game mode at the per-fold level, similarly to previous work [6], conducting a series of pairwise two-tailed Mann Whitney U tests to determine how often results from entire folds of data for each pair of game modes differed significantly in their relative predictive error.

IV. Results and Discussion

Firstly, we compare the performance of models trained using features from the Self and Group feature groups, for each target emotion. For each comparison, we consider general trends in performance data, rather than individual feature importance for brevity, with this being a focus of future discussion. Figure 3 describes the relative error seen in predictions of desire, made by models using each feature group in the three game modes. From this, it can be seen that all models struggled to attain RRSE values close to zero, although benefit in the use of a player's own facial expressions over the average of those seen across other players is clear. Models using either the Self or Group feature groups show the greatest performance when utilised in the teams game mode, when predicting desire.

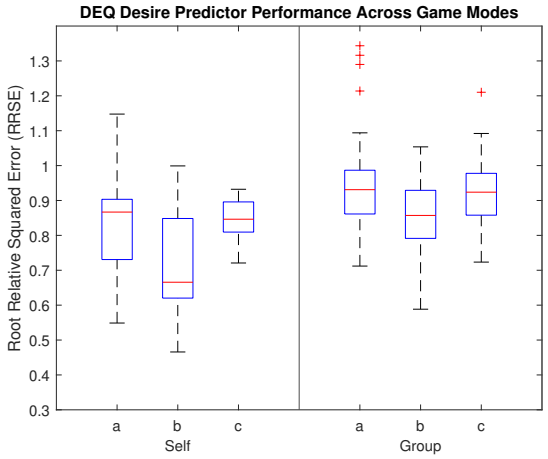

Fig. 3. A series of box plots describing the distribution of RRSE values from models utilising features representing player's own facial expressions (Self) and the average of facial expression metrics from other players in their group (Group) to predict self reported intensity of desire, as measured using the DEQ, across three different game modes in test bed game *Colour Rush*. These game modes are: competitive (a), team (b) and cooperative (c).

These trends are confirmed through statistical analysis, with the majority of pairwise tests between Self and Group models predicting responses to the same game modes showing a significant increase in error when using Group data. Similarly, pairwise tests between model performances for each pair of game modes using each feature group again showed various significant differences. Of these, results from models using the Self data (as shown in Table I) suggest that the differences seen between groups are indeed statistically significant in the majority of cases, with the team game mode providing facial expression data most usable in the prediction of the player's emotional intensity of desire, followed most commonly by the competitive game mode, then the cooperative mode.

TABLE I
P-VALUES FROM PAIRWISE MANN-WHITNEY U TESTS BETWEEN RRSES AT EACH FOLD PREDICTING DEQ DESIRE USING SELF DATA

Fold	Comp v Team	Comp v Coop	Team v Coop
1	**(+)0.008***	**(+)0.008***	**(-)0.008***
2	0.421	0.548	0.421
3	**(-)0.008***	**(-)0.032***	**(+)0.008***
4	**(+)0.008***	**(-)0.032***	**(-)0.008***
5	0.310	**(-)0.008***	**(-)0.008***

* **Significant** at $\alpha = 0.05$
(-)/(+) First game mode in pair significantly better/worse

Results in Table II again describe the success of data from the team game mode, with data from Group features more usable in this game mode than the others, in the majority of significant cases. Against expectations however, results from both Tables I and II highlight that many of the directions of effect seen in significant results across predictions of desire change from fold to fold, suggesting inconsistency in the usability of both Self and Group features in predictions for each game mode. This does explain the observed similarity seen in Figure 3 between the competitive and cooperative modes however, with varying direction of effect in significant per-fold differences resulting in similar overall distributions.

Results for the models tested in the second fold (using both Self and Group features) showed no significant difference between the three game modes, suggesting little difference between them for the second group of players. DEQ results for this group did not differ significantly from all other groups, suggesting that this consistency appropriately reflects indifference in the usability of the selected feature groups, across the three game modes, when predicting desire.

TABLE II
P-VALUES FROM PAIRWISE MANN-WHITNEY U TESTS BETWEEN RRSES AT EACH FOLD PREDICTING DEQ DESIRE USING GROUP DATA

Fold	Comp v Team	Comp v Coop	Team v Coop
1	**(+)0.008***	**(+)0.016***	**(-)0.008***
2	1.000	0.690	0.841
3	0.421	**(+)0.008***	**(+)0.016***
4	0.690	0.056	**(-)0.032***
5	**(+)0.032***	**(-)0.008***	**(-)0.008***

* **Significant** at $\alpha = 0.05$
(-)/(+) First game mode in pair significantly better/worse

In the prediction of happiness, a similar overall pattern is seen for both Self and Group models, as seen in Figure 4, with predictors for the team game mode reporting the lowest relative error values. The directionality of results in both Tables III and IV across folds suggests more consistent ordering in the usability of both player's own expression data, and that of other players in their group, across the three game modes, when predicting happiness. Specifically, for both Self and Group data, we again observe that significant results highlight the teams game mode as that for which predictive modelling is most applicable, followed by the competitive mode, and finally the cooperative game mode, although the cooperative and competitive game modes show little difference when using Self data. Of note, Group models for the competitive game mode performed more poorly than those predicting happiness in the teams mode for every fold, suggesting a highly consistent effect.

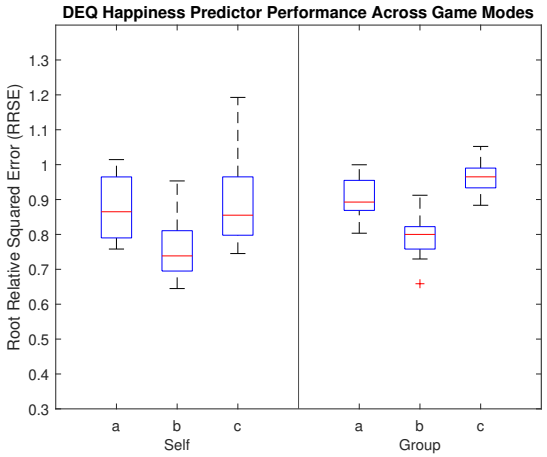

Fig. 4. A series of box plots describing the distribution of RRSE values from models utilising features representing player's own facial expressions (Self) and the average of facial expression metrics from other players in their group (Group) to predict self reported intensity of happiness, as measured using the DEQ, across three different game modes in test bed game *Colour Rush*. These game modes are: competitive (a), team (b) and cooperative (c).

Interestingly, comparisons between the Self and Group happiness performance values for each game mode highlighted

TABLE III
P-VALUES FROM PAIRWISE MANN-WHITNEY U TESTS BETWEEN RRSES AT EACH FOLD PREDICTING DEQ HAPPINESS USING SELF DATA

Fold	Comp v Team	Comp v Coop	Team v Coop
1	0.0222	0.222	0.548
2	**(+)0.008***	0.095	**(-)0.008***
3	**(+)0.016***	**(-)0.008***	**(-)0.008***
4	**(+)0.008***	1.000	**(-)0.008***
5	**(+)0.008***	0.222	**(-)0.008***

* **Significant** at $\alpha = 0.05$
(-)/(+) First game mode in pair significantly better/worse

TABLE IV
P-VALUES FROM PAIRWISE MANN-WHITNEY U TESTS BETWEEN RRSES AT EACH FOLD PREDICTING DEQ HAPPINESS USING GROUP DATA

Fold	Comp v Team	Comp v Coop	Team v Coop
1	**(+)0.008***	**(-)0.032***	**(-)0.008***
2	**(+)0.008***	0.841	**(-)0.008***
3	**(+)0.016***	**(-)0.008***	**(-)0.008***
4	**(+)0.016***	0.151	0.056
5	**(+)0.008***	**(-)0.008***	**(-)0.008***

* **Significant** at $\alpha = 0.05$
(-)/(+) First game mode in pair significantly better/worse

a difference in group three, for whom significant results suggested that (unlike all other significant differences) Group data provided models with greater predictive power than the Self data. This was seen again in predictions of relaxation, again for which the majority of significant results suggested that the Self data was more usable than the Group data, in opposition to those found for group three.

RRSE values collected from predictors of relaxation again highlight the relative success of models in the teams game mode. Figure 5 suggests a consistently greater range in performance for Self models over those using Group data, a pattern also conforming with those predicting happiness in Figure 4, potentially due to greater overfitting. Despite this, Self models perform significantly better, with models from both feature groups again suggesting the team, competitive, cooperative ordering, as confirmed by significant results in Tables V and VI. Notably, significant results describing differences between game modes for relaxation are less common than those found for happiness, but do highlight the strong difference between the team and cooperative modes most frequently.

TABLE V
P-VALUES FROM PAIRWISE MANN-WHITNEY U TESTS BETWEEN RRSES AT EACH FOLD PREDICTING DEQ RELAXATION USING SELF DATA

Fold	Comp v Team	Comp v Coop	Team v Coop
1	0.095	0.690	**(-)0.032***
2	**(+)0.008***	0.151	**(-)0.008***
3	0.151	**(-)0.008***	**(-)0.032***
4	0.690	0.151	0.151
5	1.000	**(-)0.008***	**(-)0.008***

* **Significant** at $\alpha = 0.05$
(-)/(+) First game mode in pair significantly better/worse

Considering the additional data-splits available from the teams game mode to give further insight towards RQ2, we present and analyse the relative error of models predicting emotional intensity for the three target emotions in the teams mode, using the Self, Group, Opposition and Partner feature groups.

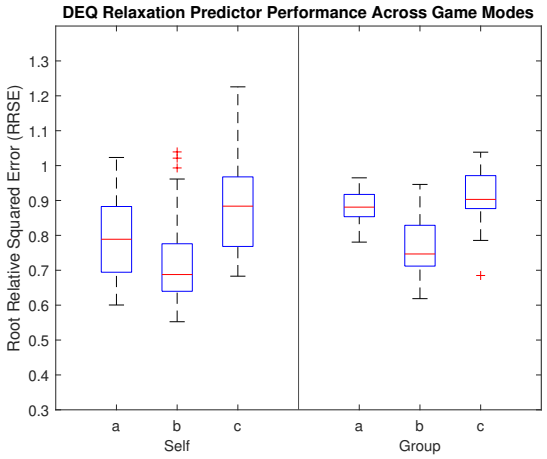

Fig. 5. A series of box plots describing the distribution of RRSE values from models utilising features representing player's own facial expressions (Self) and the average of facial expression metrics from other players in their group (Group) to predict self reported intensity of relaxation, as measured using the DEQ, across three different game modes in test bed game *Colour Rush*. These game modes are: competitive (a), team (b) and cooperative (c).

TABLE VI
P-VALUES FROM PAIRWISE MANN-WHITNEY U TESTS BETWEEN RRSES AT EACH FOLD PREDICTING DEQ RELAXATION USING GROUP DATA

Fold	Comp v Team	Comp v Coop	Team v Coop
1	**(+)0.008***	**(-)0.008***	**(-)0.008***
2	0.690	0.310	0.841
3	0.421	**(-)0.008***	**(-)0.008***
4	**(+)0.008***	**(-)0.032***	**(-)0.008***
5	**(+)0.008***	0.151	**(-)0.008***

* **Significant** at $\alpha = 0.05$
(-)/(+) First game mode in pair significantly better/worse

In the prediction of desire, shown by Figure 6, the ability of the Self data models to outperform those using inter-player data is again clear, with backing from further pairwise Mann Whitney U tests at the per-fold level confirming this in the majority of cases. Of the additional Opposition and Partner feature groups, models using the opposing players' facial expression data appear to perform least well, with the better performance of Group data models largely reflected by models using Partner features. Again this is confirmed through statistical analysis, with all significantly different sets of models at the per-fold level showing significant increase in relative error when using the Opposition features over any other feature group.

In contrast, Figure 7 shows a much greater level of success when using the Opposition feature group, over the Group and Partner features. Models using the opposing team's facial expression data performed so well in these tests that they often rivalled models using the player's own facial expression data, reporting significantly better results in two of five folds, and insignificant difference in another one, in which the two sets of models were evenly matched.

The same pattern is then seen in the prediction of relaxation, again with results from two folds reporting significant improvement over Self models when using the Opposition data, and insignificant difference at another. As can be seen in Figure 8, models predicting relaxation using Opposition data went as far as to report the lowest RRSE of the entire suite of

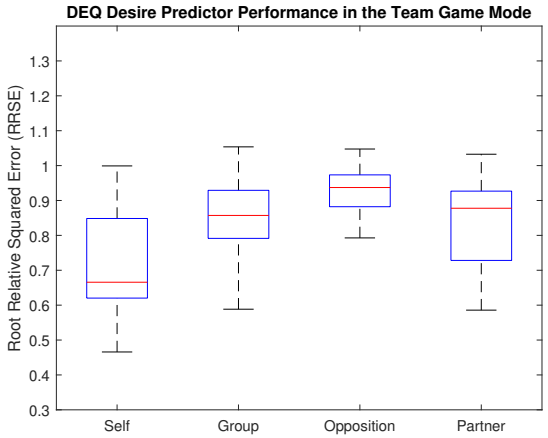

Fig. 6. A series of box plots describing the distribution of RRSE values from models utilising features representing player's own facial expressions (Self), the average of facial expression metrics from other players in their entire group (Group), the average of facial expression metrics across players on an opposing team (Opposition), and facial expressions of a player on the same team (Partner), in prediction of self reported intensity of desire, as measured using the DEQ.

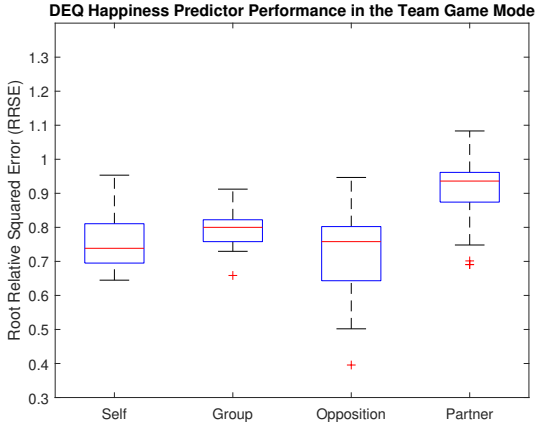

Fig. 7. A series of box plots describing the distribution of RRSE values from models utilising features representing player's own facial expressions (Self), the average of facial expression metrics from other players in their entire group (Group), the average of facial expression metrics across players on an opposing team (Opposition), and facial expressions of a player on the same team (Partner), in prediction of self reported intensity of happiness, as measured using the DEQ.

tests, across all feature groups, emotions, and game modes.

A relationship between opposing players' facial expressions and the target player's relaxation is not unexpected, especially given that this provides the most direct mapping to previous work considering tension (semantically opposed to relaxation) in one-on-one games, that showed benefit in the use of inter-player data [6]. This highlights potential in the method particularly for this style of gameplay, with players in direct opposition to the target player providing usable information rivalling that of their own expression data. The relative lack of utility seen in the Partner feature group opposed initial expectations, although this aligns with the relative error seen from predictive modelling in the cooperative game mode, and the increased performance seen in the competitive game mode. In these cases and in prediction of happiness and relaxation using the Opposition and Partner feature groups, data from

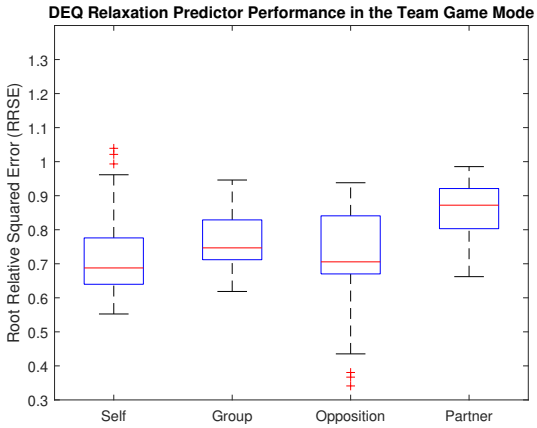

Fig. 8. A series of box plots describing the distribution of RRSE values from models utilising features representing player's own facial expressions (Self), the average of facial expression metrics from other players in their entire group (Group), the average of facial expression metrics across players on an opposing team (Opposition), and facial expressions of a player on the same team (Partner), in prediction of self reported intensity of relaxation, as measured using the DEQ.

other players working *with* the target player has been shown to be less useful than players working *against* the target player.

V. Conclusion

This study sought to respond to two research questions, targetting the potential of inter-player data across common multiplayer game modes, seeking to understand how best to utilise data collected from players in a shared experience to estimate emotional intensity in a target player. For this, we present a user study, predictive modelling and statistical analysis, and contribute to the field of human-computer interaction with findings useable in future works.

In response to RQ1, the data collected as a part of this study overwhelmingly suggests that game modes do have an effect on the usability of not only facial expression data collected from a target player for the prediction of their own self reported experience, but also data collected aggregating the facial responses of other players in the shared experience. More specifically, models using either data from the target player or data from the group, produced significantly more accurate predictions of participant's DEQ results for the desire, happiness and relaxation subscales when trained and tested on data from the team game mode, over those predicting responses collected in the competitive or cooperative game modes. The competitive and cooperative game modes often showed less significant difference in their effect on model performance, although the majority of significant results do suggest that prediction accuracy was greater when predicting for the competitive game mode. Not only does this have implications for where inter-player data may be best utilised in further work, but may provide further understanding for previous work [6], and the mixed results seen when allowing for both cooperative and competitive play [19].

In response to RQ2, we see that data from a target player provides a greater basis from which to build predictors of emotional experience than data collected from the other players in a shared experience, with this reflected across all three of

the tested game modes. When considering inter-player data-splits available to the teams game mode, our findings show that expression data averaged across the opposing team's players was usable to produce models that outperformed models trained on the player's own facial expression data in two of the five folds of cross validation, and performed equally well in a further one fold, when predicting both happiness and relaxation. This aligns with previous work showing the utility of inter-player data [6], with suggestions for further work therefore relating to its use in team based, and one-on-one games, with data from players in opposition to the target player potentially providing data useable to model their emotional response to gameplay.

Further future work may concern the collection of additional modes of input data from players, or may wish to utilise inter-player data for the prediction of any of the many measures of player experience seen in previous work.

Limitations to our work largely concern the sample size at which our data was collected, a pressure for all work focussed on multiplayer games studies. Efforts were made to utilise as much data as possible in the selection of features and training, utilising nested cross validation to prevent data leakage, although a greater sample size would have proved useful in ensuring greater repeatability across player groups, and provided a greater range of demographics to lend further usability to our results. A predominantly male participant group may have impacted findings, with previous work suggesting that competitive game modes result in a greater level of emotional expression from male participants [10] potentially accounting for the greater utility seen in data from competitive gameplay seen in this study. Given a greater number of participants, further work may also consider exploration into the number of players per group in group-based games, with potential in testing the effect of team sizes on the positive results found in this study.

REFERENCES

[1] D. Villani, C. Carissoli, S. Triberti, A. Marchetti, G. Gilli, and G. Riva, "Videogames for emotion regulation: a systematic review," *Games for health journal*, vol. 7, no. 2, pp. 85–99, 2018.

[2] C. Von Scheve and S. Ismer, "Towards a theory of collective emotions," *Emotion review*, vol. 5, no. 4, pp. 406–413, 2013.

[3] M. Wróbel, "I can see that you're happy but you're not my friend: Relationship closeness and affect contagion," *Journal of Social and Personal Relationships*, vol. 35, no. 10, pp. 1301–1318, 2018. [Online]. Available: https://doi.org/10.1177/0265407517710820

[4] J. Quan, Y. Miyake, and T. Nozawa, "Incorporating interpersonal synchronization features for automatic emotion recognition from visual and audio data during communication," *Sensors*, vol. 21, no. 16, p. 5317, 2021.

[5] G. Pereira, J. Dimas, R. Prada, P. A. Santos, and A. Paiva, "A game prototype with emotional contagion," in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 315–316.

[6] P. Mavromoustakos-Blom, D. Melhart, A. Liapis, G. N. Yannakakis, S. Bakkes, and P. Spronck, "Multiplayer tension in the wild: A hearthstone case," in *Proceedings of the 18th International Conference on the Foundations of Digital Games*, 2023, pp. 1–9.

[7] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 653–656.

[8] K. Pinitas, N. Rasajski, M. Barthet, M. Kaselimi, K. Makantasis, A. Liapis, and G. N. Yannakakis, "Varying the context to advance affect modelling: A study on game engagement prediction," in *Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2024.

[9] G. Chanel, J. M. Kivikangas, and N. Ravaja, "Physiological compliance for social gaming analysis: Cooperative versus competitive play," *Interacting with Computers*, vol. 24, no. 4, pp. 306–316, 05 2012. [Online]. Available: https://doi.org/10.1016/j.intcom.2012.04.012

[10] J. M. Kivikangas, J. Kätsyri, S. Järvelä, and N. Ravaja, "Gender differences in emotional responses to cooperative and competitive game play," *PLOS ONE*, vol. 9, no. 7, pp. 1–16, 07 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0100318

[11] P. A. Nogueira, R. Aguiar, R. Rodrigues, and E. Oliveira, "Computational models of players' physiological-based emotional reactions: A digital games case study," in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 3, 2014, pp. 278–285.

[12] G. Chanel and P. Lopes, "User evaluation of affective dynamic difficulty adjustment based on physiological deep learning," in *Augmented Cognition. Theoretical and Technological Approaches*, D. D. Schmorrow and C. M. Fidopiastis, Eds. Cham: Springer International Publishing, 2020, pp. 3–23.

[13] C. T. Tan, S. Bakkes, and Y. Pisan, "Inferring player experiences using facial expressions analysis," in *Proceedings of the 2014 Conference on Interactive Entertainment*, 2014, pp. 1–8.

[14] R. Somarathna and G. Mohammadi, "Towards understanding player experience in virtual reality games through physiological computing," in *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 2024, pp. 405–408.

[15] C. Harmon-Jones, B. Bastian, and E. Harmon-Jones, "The discrete emotions questionnaire: A new tool for measuring state self-reported emotions," *PLOS ONE*, vol. 11, no. 8, pp. 1–25, 08 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0159915

[16] D. Bonk and J. Kim, "Factorial and construct validity of the discrete emotions questionnaire for videogames (deq-vg)," *Entertainment Computing*, vol. 42, pp. 100 488, 8, 2022.

[17] E. S. Siqueira, M. C. Fleury, M. V. Lamar, A. Drachen, C. D. Castanho, and R. P. Jacobi, "An automated approach to estimate player experience in game events from psychophysiological data," *Multimedia Tools and Applications*, vol. 82, no. 13, pp. 19 189–19 220, 2023.

[18] Unity Technologies, "Unity," 2005.

[19] A. Brooke, M. Crossley, H. Lloyd, and S. Cunningham, "Towards predicting player experience as discrete emotion intensity using gameplay and visual data in a multiplayer game," 2024. [Online]. Available: https://doi.org/10.36227/techrxiv.172651567.77028268/v1

[20] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE transactions on cybernetics*, vol. 43, no. 6, pp. 1519–1531, 2013.

[21] P. M. Blom, S. Bakkes, and P. Spronck, "Modeling and adjusting in-game difficulty based on facial expression analysis," *Entertainment Computing*, vol. 31, p. 100307, 2019.

[22] P. Mavromoustakos-Blom, M. Kosa, S. Bakkes, and P. Spronck, "Correlating facial expressions and subjective player experiences in competitive hearthstone," in *Proceedings of the 16th International Conference on the Foundations of Digital Games*, 2021, pp. 1–5.

[23] E. Hatfield, "Primitive emotional contagion," *Review of personality and social psychology: Emotion and social behavior/Sage*, 1992.

[24] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[25] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[26] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the national academy of sciences*, vol. 111, no. 15, 2014.

[27] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling player experience for content creation," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 2, no. 1, pp. 54–67, 2010.

[28] H. P. Martínez and G. N. Yannakakis, "Genetic search feature selection for affective modeling: a case study on reported preferences," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, 2010, pp. 15–20.