DM

# Tailoring a Psychophysiologically Driven Rating System

MASTER DISSERTATION

**Harryharasuthan Vasantharajah**
MASTER IN COMPUTER ENGINEERING

UNIVERSIDADE da MADEIRA
A Nossa Universidade
www.uma.pt

September | 2019

# Tailoring a Psychophysiologically Driven Rating System

MASTER DISSERTATION

**Harryharasuthan Vasantharajah**
MASTER IN COMPUTER ENGINEERING

SUPERVISOR
Sergi Bermúdez I Badia


# TAILORING A PSYCHOPHYSIOLOGICALLY DRIVEN RATING SYSTEM

**Harryharasuthan Vasantharajah**
**B.Sc. (Hons)**

Supervised by

Sergi Bermúdez I Badia, Ph.D.

Submitted in fulfillment of the requirements for the degree of
Masters in Computer Engineering.

Faculty of Exact Sciences and Engineering
University of Madeira
2019

# Abstract

Humans have always been interested in ways to measure and compare their performances to establish who is best at a particular activity. The first Olympic Games, for instance, were carried out in 776 BC, and it was a defining moment in history where ranking based competitive activities managed to reach the general populous. Every competition must face the issue of how to evaluate and rank competitors, and often rules are required to account for many different aspects such as variations in conditions, the ability to cheat, and, of course, the value of entertainment. Nowadays, measurements are performed out through various rating systems, which considers the outcomes of the activity to rate the participants. However, they do not seem to address the psychological aspects of an individual in a competition.

This dissertation employs several psychophysiological assessment instruments intending to facilitate the acquisition of skill level rating in competitive gaming. To do so, an exergame that uses non-conventional inputs, such as body tracking to prevent input biases, was developed. The sample size of this study is ten, and the participants were put on a round-robin tournament to provide equal intervals between games for each player.

After analyzing the outcome of the competition, it revealed some critical insights on the psychophysiological instruments; Especially the significance of Flow in terms of the prolificacy of a player. Although the findings did not provide an alternative for the traditional rating systems, it shows the importance of considering other aspects of the competition, such as psychophysiological metrics to fine-tune the rating. These potentially reveal more in-depth insight into the competition in comparison to just the binary outcome.

***Keywords:***

ELO, Exergame, Game Design, Machine Learning, Psychophysiology, Rating Systems

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| AI | Artificial Intelligence |
| ECG | Electrocardiography |
| EEG | Electroencephalography |
| EMG | Electromyography |
| FOV | Field of View |
| FSM | Finite State Machine |
| GP | Games Played |
| HR | Heart Rate |
| HRV | Heart Rate Variability |
| IR | Infrared |
| LR | Linear Regression |
| ML | Machine Learning |
| PA | Positive Affect |
| PANAS | Positive and Negative Affect Schedule |
| PA | Points Against |
| PD | Points Difference |
| PF | Points For |
| PPG | Photoplethysmography |
| PPO | Proximal Policy Optimization |
| PTS | Points |
| NA | Negative Affect |
| RD | Rating Deviation |
| RL | Reinforcement Learning |
| RQ | Research Question |
| SAM | Self-Assessment Manikin |
| SD | Standard Deviation |
| TLX | Task Load Index |
| TTT | TrueSkill Through Time |
| UGE | Unity Game Engine |
| VR | Virtual Reality |

# Acknowledgments

First of all, I would like to convey my sincere gratitude to all the extraordinary people who have been generous enough to dedicate some of their time, knowledge, and patience during my master's degree.

I am tremendously thankful to my supervisor Dr. Sergi Bermudez I Badia, for his guidance, constant encouragement, valuable suggestions, critical comments, and perfectionism during my thesis. I would also like to thank my friend Eduardo Gomes, for dedicating his time and sharing his knowledge in machine learning aspects and data science as well as leading thoughtful discussions throughout my dissertation work.

I am incredibly grateful to my mentors Dr. Filipe Quintal and Dr. Lucas Pereira, for allowing me to commence my career in Madeira. My dearest friend and colleague, Deise Faria, for being there by my side for a shoulder to lean on. Alexandra Mendes, Angela Barbosa, and Helena Barbosa for their continuous encouragement.

I am enormously grateful for former NeuroRehabilitation Lab members, Dr. Athanasios Vourvopoulos and Dr. John Muñoz, and current members Teresa Paulino, Yuri Almeida, Afonso Gonçalves and the rest of the relentlessly hardworking team. Special thanks to all participants who spent their valuable time to contribute to the outcome of this dissertation.

I would also like to thank Carina for her love, encouragement, humor, and energy that has given me strength in even in the darkest hours of my life and my family for allowing me to pursue my dreams and goals with self-responsibility and independence since my childhood.

Finally, for everyone, I have forgotten to mention, and everyone who was part of my life; every one of you made this dissertation possible.

My heartfelt thank you!

# Chapter 1: Introduction

This chapter outlines the introductory background (section 1.1) and motivation (section 1.2) of the research that has been conducted and its objectives (section 1.3). Section 1.4 describes the significance and scope of this dissertation as well as provides definitions of terms used. Finally, section 1.5 includes an outline of the remaining chapters of the thesis.

## 1.1 BACKGROUND

Humans have always been interested in ways to measure and compare their performances to establish who is best at a particular activity. The first Olympic Games, for instance, were carried out in 776 BC, and it was a defining moment in history where ranking based competitive activities managed to reach the general populous. Nowadays, competitions are carried out for almost any discipline one can compete in, including sports, games or mental challenges and some competitions, such as the football world-cup, attracting a vast number of spectators. In the past, game competitions were generally not as popular as sports, although there were trends in countries like China and South Korea. At present, game competitions have reinvigorated themselves with the title of eSports[1] and have the capability to fill stadia daily [2].

Every competition must face the issue of how to evaluate and rank competitors, and often rules are required to account for many different aspects such as variations in conditions, the ability to cheat, and, of course, the value of entertainment. Evaluating prolificacy in gaming by means of measuring and logging physiological and psychophysiological responses such as Heart Rate (HR) metrics and Electrodermal Activity/Galvanic Skin Response (GSR), FLOW, Perceived Workload, etc. of the players during a gameplay session is a prominent method in the field of game user research. Besides, these methods are widely used in assessing the expertise of pilots[3], astronauts[4], surgeons[5], and soldiers[6] in simulations and virtual environments.

For years, these techniques have been out of reach for many researchers due to the underlying limitation in computational power, lack of exposure to psychophysiological instruments, and the cost of physiological sensors along with the complexity of implementing them (e.g., Electrocardiography- ECG) in a non-intrusive manner. The recent boom in computational power, data science, and machine learning as well as affordable composite sensors such as, smartwatches which uses Photoplethysmography (PPG) based sensors.

## 1.2 MOTIVATION

Being an avid competitive gamer myself, from my childhood, I always had a fascination for competitive gaming and the sense of climbing the ladder of leader boards in video games. Furthermore, I have worked with several researchers of Game User Research, Exergames, and other Serious Games at the NeuroRehabilitation Lab of Madeira Interactive Technologies Institute and the University of Madeira on various physiological sensors and psychophysiological assessment instruments.

One of the crucial catalyst to pursue this topic was my work with NeuroRehabilitation Lab members John Muñoz and Teresa Paulino on developing and co-authoring Android-based framework component for wearables such as smartwatches to extract HR (Heart Rate) on-demand to be applied in mobile VR (Virtual Reality) environments [7]. Besides, further collaboration with the authors lead to development and integration other proprietary sensors (such as CardioBAN, Polar H10 as well as the Myo Armband) to their work on The Biocybernetic Loop Engine [8], which allowed me to grasp the fundamentals of the field, as well as feasibility of technologies that can be pragmatically applied on my research.

Thus, I embarked on a journey to disentangle different prolificacy of players and tailor a rating system in competitive gaming using psychophysiological metrics.

## 1.3 OBJECTIVES

This dissertation employs several psychophysiological assessment instruments as well as cost-effective wearables (that utilizes PPG sensors) as an alternative to the not so contemporary intrusive sensors, intending to facilitate the acquisition of player skill level rating in conventional gameplay scenarios. Mainly focusing on investigating the role of psychophysiological states such as flow, challenge, and dominance.

This work hypothesizes that by taking advantage of these tools, it is possible to discriminate the different prolificacy (skill level or expertise) of players in competitive gaming not only by the outcome of the game but also by behavior of the players in both in-game (during the activity) and off game (prior and post-activity). Thus, this dissertation will be focusing on interrelating the player's relative skill level rating to their psycho-physiological metrics during a particular gameplay session.

Thus, this dissertation aims to explore the following research questions:

***RQ1. How relative competitive skill affects the subjective experience of players in gaming?***

***RQ2. How subjective experience influences the absolute competitive skill of players in gaming?***

***RQ3. Can the addition of subjective experience be beneficial for the traditional skill rating system?***

The outcomes of these questions and the thesis itself can be useful for game user researches, who are focusing on competitive game rating aspects as well as for game designers who may use the results to build upon the enjoyability of competitive gaming by leveraging both the qualitative and quantitative results.

## 1.4 SIGNIFICANCE, SCOPE, AND DEFINITIONS

Prolificacy (expertise at its highest levels) has been studied from several areas and fields. These include both academics (e.g., physics, chemistry, and mathematics) and non-academic domains (e.g., chess, typing, solving a Rubik's cube and restaurant ordering)[9]. However perceived, representations of the expertise describe characteristics of prolificacy in individual terms. Initial features, including automaticity, speed of processing information, visualization, etc. have all been used to explain how prolific individual perform within their specific domain.

Furthermore, attributes like age have been proposed to advocate how these experts advance within a specific domain. Notably, commitment at a young age to a field correlates to higher levels of prolificacy in that area [10]. The literature on prolificacy, however, is not necessarily formulated to describe the development of skill in extremely dynamic, immersive settings. For example, current digital environments are also highly collaborative and social [11]. Except for the mentoring/guidance role in deliberate practice, the literature on high levels of expertise seldom tackles the social aspects of learning [10], [12].

According to Murphy and Alexander [13], prolificacy is centered fundamentally on the maturity of domain knowledge. Based on the activities described at the beginning of this section (section 1.4), this would be no different in playing competitive videogames where players spend a tremendous amount of time honing skills, researching information, and put what they have learned to practice. As with all hyper-environments, users (i.e., gamers) are responsible for efficiently and effectively finding and evaluating information, apprehending information across multiple modalities simultaneously, and orchestrating dynamic strategies that facilitate learning in these complex environments[14]. However, domain knowledge (i.e., game content, mechanics, etc.) and the means to acquire it are not the only areas in

which gamers need to excel. Concerning competitive videogames, successful players must also master the technology.

Mastering technology is tied to simple tasks such as playing the game to more complex tasks associated with optimizing the game experience. Some players spend hours perfecting a simple action or honing niche game mechanics using unconventional methods to have the edge over their opponents. It follows that developing expertise in competitive video games involves interaction with and proficiency in several distinct areas.

Hence, this dissertation will have delimited its scope to one versus one competitive environment where the game uses non-conventional yet intuitive inputs for the players to avoid experience bias that was discussed earlier in this section (section 1.4). Besides, this work incorporates mild exergaming aspect along with gradually increasing pace [15] to the gameplay to hasten and amplify the process of identifying key indicators such as changes in HR as well as other psychophysiological metrics such as perceived workload, exertion, anxiety, arousal, dominance, challenge, and flow.

## 1.5 THESIS OUTLINE

***Chapter 1 - Introduction***

This chapter began with the detailed background and motivation of this work, as well as the objectives of this dissertation. Finally, it stated the significance of the thesis, along with the research scope and definitions.

***Chapter 2 - Literature Review***

Introduces the reader to the historical background of performance rating systems and continues onto a detailed chronological review of prominent rating systems.

***Chapter 3 - Research, Design & Development***

Explores various instruments that are to be utilized for this study and moves onto technological choices. This chapter closes with the design and development of solutions that aided in the design of the study.

***Chapter 4 - Methodology***

Focuses on the procedure of the research, commencing from the competition format, reasoning behind the choice of participants, the number of sessions, and the step-by-step process of the study in detail.

***Chapter 5 - Findings***

Presents the outcome of the study in stages based on various instruments utilized from the study design to provide a general overview of the results.

***Chapter 6 - Conclusions***

It provides insights into the implications and interpretation with reference it the literature and states the author's take on the findings of the dissertation along with its potential limitations as well as the conclusions.

# Chapter 2: Literature Review

This chapter begins with a historical background (section 2.1) and reviews literature on the following prominent rating systems: Ingo (section 2.2), Elo (section 2.3) Glicko (section 2.4), Edo (section 2.5) and TrueSkill (section 2.6). Section 2.7 synopsizes all the above rating systems by history, comparison, advantages, and limitations.

## 2.1 HISTORICAL BACKGROUND OF PERFORMANCE RATING SYSTEMS

Numerous ranking systems instigated with modern chess ranking systems as early as the 1930s [16], and more recently are applied widely in online competitive gaming rating, for example, in Counter-Strike, Dota2 [17], [18] and League of Legends [19], [20] as well in gaming systems such as the Microsoft Xbox entertainment system. Competitive games such as chess tend to use skill rating systems for several practical purposes: (a) to qualify candidates for elite tournaments, (b) to pair candidates of similar abilities for tournaments, and (c) to monitor candidates' progress [21].

In general, rating systems are designed to provide information about players' skill development by combining data from a new game outcome with players' skills, as demonstrated from previous games. These systems aim to provide information about a player's strength at any time. Systems such as ELO, update players' strength estimates after each game, whereas others such as TrueSkill update information after a series of games. These systems were initially developed to rank two-player games, and in more recent years, ranking systems have been further developed to rank players in multiplayer games.

The purpose of this section is to provide information about some of the most well-known existing ranking systems and then summarize, compare, and contrast some of the most renowned systems for two-player or multiplayer games. They are the Ingo, Elo, Glicko, Edo, and TrueSkill ranking systems.

## 2.2 INGO RATING SYSTEM

One of the first ranking systems to produce numerical ratings, the Ingo system was developed by Anton Hoesslinger in 1948 and used by the German Chess Federation [16]. Over the following decade, many versions of this system were developed and used in different national chess tournaments. The Ingo system was used for paired comparisons. Unlike contemporary ranking systems such as ELO, TrueSkill, etc., the Ingo system associates better performance with lower scores.

The Ingo system is considered a simple one, with little basis in statistical ratings. A player's ranking is based on the performance of the average player. In particular, the average rating of the players in a competition is calculated. Also, the player's score in percentage points is calculated.

Equation 1. Calculating INGO

$$R = O - (W - 50)$$

R is the player's new rating, O is the arithmetic average of the ratings of the player's opponents, and W is the player's win ratio express as a percentage. If a player's percentage score is average (50%), then the player's rating score is the average rating score; if the player's percentage score is above 50%, then the player receives the average score plus 10 points for each percentage point above 50%. Similarly, if the player's percentage score is below 50%, then the player receives the average score minus 10 points for each percentage point below 50%.

For example, if the average rating score in the competition is 1,500 and the percentage score of a player is 23%, then this score is 27 percentage points below average, so the new rating score of the player is 1,500 − (10 * 27) = 1,230.

## 2.3 ELO RATING SYSTEM

The Elo system was developed by Arpad Elo in 1959 and adopted by the World Chess Federation in 1970 [22]. It is probably the most widely used system in competitive games such as chess. Like the Ingo system, the Elo system is a ranking system for two-player games. However, the Elo system is based on a model with a considerably more statistical foundation. The Elo system assigns a number between 0 and 3,000 that changes over time based on the outcomes of tournament games. Unlike the Ingo system, in the Elo system, a higher score indicates better performance. Thus, a player with a higher rating is expected to win more often than a player with a lower rating. Based on the game outcomes, the player's rating may be increased or decreased.

The primary assumption of the Elo system is that each player is associated with a current strength, and a rating estimates this strength. The Elo system associates game results in latent variables that represent the ability of each player. The Elo system uses the Thurstone-Mosteller model to estimate the probability of individual game outcomes based on the assumption that the player's chess performance in each game is a random variable that is typically distributed. It is assumed that the actual ability of each player is the mean of that player's performance. Performance is measured by wins, losses, and draws.

The assumption that a player's performance is normally distributed raises some concerns. Some statistical tests have indicated that this assumption does not accurately represent the actual results, especially for weaker players, who have higher chances to win than Elo predicts. For this reason, some chess sites use a logistic distribution. The logistic distribution version of the system goes back to Zermelo [23], who developed a model for paired comparisons that later became known as the Bradley-Terry model [24], [25]. The Bradley-Terry model is an approach to ranking n individuals by comparing two at a time.

One of the greatest assets of the Elo system in terms of usability is its linear approximation. The linearization of this model makes it attractive to users due to its simplicity. If players win more games than expected, their ratings will increase. Similarly, if players lose more games than expected, their ratings will decrease. However, the adjustment is assumed to be linearly related to the number of wins/losses by which the players differ from their expected number of wins/losses.

Furthermore, players' performance ratings are a function of the opponent rating and a linear adjustment to the amount by which they overperform or underperform their expected values. All things being equal, when players' actual scores are less than the expected values, their ratings are adjusted downward. On the other hand, if their actual scores are higher than their expected scores, the ratings are adjusted upward. The rating update for each player can be performed after each game or after a defined rating period.

Although the linear nature of this model makes it simple, advances in technology have made it obsolete. One of the limitations of the simplicity of the Elo model is that more efficient estimation models are becoming more attractive. Another limitation of the Elo model is that it uses a player's most recent rating as the current one, even if the player has not competed for a long time.

Nevertheless, the Elo rating system can be used not just for rating players. It has been used for rating patterns in the game of Go [26], eliciting user preferences [27], assessing security and vulnerability risks [28], ranking posts in online forums [29], choosing the efficient layout to reduce fabric waste in clothing industry [30], as well as a plethora of application in the field of soft biometrics such as human description identification [31]–[33] (body, posture, and movement) and human facial identification [34], [35]. Recently several animal behavioral scientists also used Elo to estimate social dominance strengths and of animals in the wild [36]–[38].

## 2.4 GLICKO RATING SYSTEM

Glickman developed the Glicko system in 1995 [39]. Like the Ingo and Elo systems, the Glicko system is designed for two-player games. This model is an extension of the Elo system and was developed in an attempt to address and improve the parameter estimates by incorporating a variability factor. The Glicko system computes the rating similarly to the Elo system, but it also incorporates the reliability of a player's rating. The reliability of a rating is called the rating deviation (RD), which is a standard deviation that measures the uncertainty of the rating. For example, a player who did not play for a long time and had just one game may have a high RD. A player who competes very often may have a low RD. The rationale is that the system can gather more information about the skill of the player who competes more often, and therefore the rating is more precise than that of a player who competes less often. Because the Glicko system provides both a rating and an RD, it may be more informative to describe players' skills as a confidence interval. For example, a 95% confident interval is calculated as Rating ± 2 * RD [21], [39].

According to Weng and Lin [40], the Glicko system was the first to use the Bayesian ranking system. It is assumed that the skill of the players follows a Gaussian distribution. The Glicko system applies the Zermelo model [23], better known as the Bradley-Terry model [24], [25]. As mentioned earlier, the Bradley-Terry model is an approach to rank n individuals by comparing two at a time. The Glicko system updates the skill of the players after each rating period. For better estimates, the number of games in each rating period is between 5 and 10 games for each player [40]. A drawback of the original Glicko system (Glicko-1) is that it may not capture the exact change in skills for players who frequently compete because the RD is small for players who compete very often. As a result, the rating for these players may not change accurately [39].

In addition to the Glicko-1 system, Glickman developed the Glicko-2 system. The Glicko-2 adds rating volatility to the rating and RD. The rating volatility index is the degree of expected fluctuation in a player's rating. The volatility measurement is low when a player has consistent results, and it is high when a player has an inconsistent performance. As with the Glicko1 system, results for the Glicko-2 system are updated after a rating period. Like the Glicko-1 system, Glicko-2 performs best when rating periods consist of 5 to 10 games for each player. It should be noted that the rating outcomes based on the Glicko-2 system are very similar to the ones from the Glicko-1 system because the results do not incorporate any evidence of the volatility index [41].

## 2.5 EDO RATING SYSTEM

The Edo rating system has been developed and maintained by Rod Edwards since 2004 [42]. Similar to the systems discussed above, Edo is a rating system for paired comparisons. Also, like the Glicko system, Edo is based on the Bradley-Terry model [23], [24]. Its mean rating is adjusted to roughly 1,500 with a standard deviation of around 300.

What makes the Edo system distinctive is that during the rating/estimation, the system treats the same player at two different years as two different players. The rating of players who participated in matches in two different years is then computed as a weighted rating between the two years as if the players had played against themselves in those years. The weight is set up around 50%. A weight higher or lower than 50% can compensate for inflation or deflation of the rating from time to time (e.g., due to a player's skill increase). Also, according to Edwards [42], more self-matches of the same player result in a more stable rating of the player, whereas fewer such games mean that the player's rating is more the result of current performance.

Because at the end of the 20th century, more local tournaments with players at the lower end of the rating skill were included compared to earlier times, there is a tendency during modeling for estimation to be pulled down when more local tournaments are recorded. The second distinctive factor of the Edo system is that an adjustment is made to account for this situation: Players with ratings higher than 1,500 are marked down, while players with ratings lower than 1,500 are elevated. After this adjustment, the maintained result is similar to that of the Elo system.

Besides, Edwards [42] also claimed that the Edo system has advantages in measuring uncertainties when compared to the Glicko system. For example, when a small group of players has played against one another but not often against players outside of the group, the Edo system has "some links" [42] to the main group under this situation.

However, it is unclear how these links are maintained and estimated. Furthermore, although this model considers information for the same player at different times and provides variance of the player's skill, it does not offer posterior distributions, is not a full Bayesian model, and does not model draws [43].

## 2.6 THE TRUESKILL RATING SYSTEM

The TrueSkill model was developed by Microsoft Research [44] and maybe viewed as a generalization of the Elo system to multiplayer games. The TrueSkill ranking system is used for Microsoft's Xbox online games, and in general, it is used to rank players for video games with more than two players and/or teams per match in competitive games. The simplest scenario for TrueSkill is the same as the one described in the Elo and the Glicko systems for two players competing against each other. However, the TrueSkill model was reported to provide more accurate estimates in predicting game outcomes and in matching players compared to the Elo system [44].

The TrueSkill system uses Bayesian approximation estimation [45], [46], which allows for instant ranking updates of players and/or teams after each game. In a game, each player is assumed to have a prior skill with a mean and a standard deviation, and a Gaussian distribution is assumed. In Xbox Live, a previous skill with a mean of *25* and a variance of *(25/3)2* is used for the initial run. The performance of players in a game has a mean around their estimated skill with a standard deviation. The performance of a team is the sum of each member's performance. Each team's performance is then compared to decide team ranking. Draws (players with equal ranks of performance) are allowed in the TrueSkill ranking system.

If the difference between two teams in terms of their performance is less than a draw margin, these two teams are ranked at the same level. The draw margin can be narrow or broad, depending on the needs of the estimation. A small margin should be used when individuals'/teams' skills are relatively close, and fewer ties must be observed in the ranking. On the other hand, a wide margin should be used when ranking is more entertaining and low stakes. Posterior estimation of each player's skill is then used as a prior for ranking estimate of the player's next game. The estimation algorithm of TrueSkill uses approximate message passing—a Bayesian approximation

method [45], [46]. It is reported that convergence is fast; thus, instantaneous ranking is possible [40].

The initial TrueSkill rating system ranks game players at a particular time point *(t)* by updating their earlier rankings *(t-1)* as the prior and always estimates players' rankings forward through time. Dangauthier et al. [43] extended TrueSkill to assess players' skills not only forward through time but also backward. They called this extension TrueSkill Through Time (TTT) or TTT-D when the estimation of an additional draw margin parameter discussed earlier is included. Under TTT, for example, if Player A beats Player B, and then later, Player B beats a strong Player C, TTT and TTT-D can adjust Player A's ranking by going backward in the estimation. However, the original TrueSkill rating system is not able to make the backward adjustment for Player A in this case. However, a longer estimation time is required and inevitable because there are more steps in the algorithm when estimation goes forward or backward to consider the ranking of players who were rated previously, and adjustment is needed when new players are lined up to be ranked.

An essential feature of the TrueSkill ranking system is player matchmaking [44]. For players to have a competitive and enjoyable gaming experience, the skills of competitors have to be close. TrueSkill can match online players based on their estimated skills. There are two scenarios: games of individuals and games of teams. In a multiplayer (nonteam) competition, a simple criterion used for matchmaking is to ensure that the players' highest and lowest ratings in a game do not go above a predetermined rating difference. In a multiteam match, a team member's ranking is estimated with all the other players to get a pairwise rating. For each player, relative pair standings are then averaged as the player's ranking. The criterion for multiteam game matchmaking is to have about the same number of players on each team and also for all team players across teams to have similar skill levels.

In addition to the original TrueSkill model, several TrueSkill variant models have been used for online data: multilabel classification [47] and Web commercial click rate prediction for Microsoft's Bing search engine [48].

## 2.7 SUMMARY

### 2.7.1 Ingo System Overview

***History***

A first chess rating system developed in 1948 by Anton Hoesslinger and adopted by the German Chess Federation. In the decade after its development, several versions of this system were developed.

***Comparison***

Has a little basis in statistical theory. Calculates player's ranking based on the performance of the average player. Lower scores indicate higher performance.

***Advantages***

A straightforward model for implementing and the ratings were consistent with the subjective ranking of chess players.

***Limitations***

A player could lose in every game and still gain rating points.

### 2.7.2 Elo System Overview

*History*

The most widely used system in competitive games. Developed in 1950 by Arpad Elo as an improved rating system over the Ingo system and adopted by the World Chess Federation in 1970.

*Comparison*

It is based on a model with a considerably more statistical foundation compared to the Ingo system. The performance rating of a player is a function of the opponent rating and a linear adjustment to the amount by which a player overperformed or underperformed that player's expected value. All things being equal, when a player's actual score is less than that player's expected value, the rating is adjusted downward. On the other hand, if the actual score is higher than that player's expected score, the rating is adjusted upward. Higher scores indicate better performance. For example, when two players compete, the system predicts that the player with a higher rating is expected to win more often than the player with a lower rating. It uses two different distributions and assumes that players' performance distribution follows either a normal or a logistic distribution.

*Advantages*

Applies the Thurstone-Mosteller model, and the range of the rating scores is between 0 and 3,000.

*Limitations*

Uses player's most recent ratings as the current rating, even if the player has not competed for a very long time

### 2.7.3 Glicko System Overview

***History***

Developed by Glickman in 1999 as an extension of the Elo system. One may think of the Elo system as a particular case of the Glicko system because it not only computes the player's rating but incorporates the reliability of the player's rating called rating deviation (RD).

***Comparison***

Uses a Bayesian ranking system that applies the Bradley-Terry model based on the assumption that the player's skill distribution follows a Gaussian distribution. The rating update for each player can be computed after each rating period.

***Advantages***

Attempts to improve the parameter estimates by incorporating the rating deviation (RD).

***Limitations***

It may not capture the exact change in skills for players who frequently compete because the RD is small for them, which leads to inaccuracies in ratings.

### 2.7.4 Edo System Overview

*History*

Developed by Rod Edwards, which treats the same player at two different years as two different players.

*Comparison*

It is based on the Bradley-Terry model and provides variance of the player's skill. An adjustment is made to maintain the rating with a mean of 1,500 and a standard deviation of 300 because more players at the lower end of the rating were included at the end of the 19th century.

*Advantages*

It is claimed to estimate isolated players better than Glicko [42].

*Limitations*

It is not a full Bayesian model, and it does not provide a posterior distribution and provides ratings only up until 1910. Although the Edo rating system was developed in 2004, it used significantly older data for rating.

### 2.7.5 TrueSkill System Overview

*History*

Developed by Microsoft Research in 2007 and adopted by Xbox Live game, Microsoft's Bing search engine, and Internet information multilabel classification. A player's skill and performance are updated after each game.

*Comparison*

TrueSkill Ranking system matches players or teams of players with similar skills by utilizing Bayesian approximation (assuming it is a Gaussian distribution) factor graphs and a sum-product algorithm to allows instantaneous ranking updates. Each team's performance is the sum of its team members' performance. Draws are allowed in the system, and the margin of the draw can be adjusted. Thus, allowing players to experience enjoyable gameplay by matchmaking equally skilled opponents.

*Advantages*

Skill level estimation is instantaneous. It is reported that the estimation of TrueSkill is more precise than that of Elo [27], [49].

*Limitations*

The bayesian approximation is a compromise among estimation precision, speed, and resources. The system will need some initial infrastructure building.

# Chapter 3: Research, Design & Development

This chapter describes the research, design, and development methods adopted by this dissertation to address the research questions stated in section 1.3 of Chapter 1: *RQ1. How relative competitive skill affects the subjective experience of players in gaming? RQ2. How subjective experience influences the absolute competitive skill of players in gaming? RQ3. Can the addition of subjective experience be beneficial for the traditional skill rating system?* Section 3.1 discusses in detail the instruments to be used in the study. Section 3.2 elaborates the activity (exergame) to be performed for the study, and section 3.3 explores the choice of appropriate technology to accommodate the stud. Section 3.4 outlines the design and development process of the game as well as all related solutions relevant to this dissertation.

## 3.1 INSTRUMENTS

### 3.1.1 ELO

According to Arpad Elo's original work, the rating system implicitly characterizes the probability of winning against other players, whose Elo rating is known to us. The table below (Table 1) summarizes the fact that this probability relies solely on the rating discrepancy amongst the two players.

| Rating Difference | Winning Probability |
|---|---|
| **+400** | .919 |
| **+300** | .853 |
| **+200** | .758 |
| **+100** | .637 |
| **+50** | .569 |
| **0** | .500 |
| **-50** | .431 |
| **-100** | .363 |
| **-200** | .242 |
| **-300** | .147 |
| **-400** | .081 |

Table 1. Elo rating difference and winning probability

The fundamentals of the Elo rating system can be summarized as follows. For each player $i$ we have a rating estimate $\theta_i$. Let $R_{ij} \in \{0, 1\}$ be the results of a match amongst players $i, j$. The predicted probability that the player $i$ wins is represented by the logistic function (produces values similar to the one of Table 1 and generates a sigmoid curve identical to Figure 1) with respect to the difference of estimated ratings:

Equation 2. Predicted probability

$$P(R_{ij} = 1) = 1/(1 + e^{-(\theta i - \theta j)})$$

Based on the outcome of a match, the rating estimates are refreshed using the following update rule (*K* is a constant denoting sensitivity of the estimate to the last attempt):

Equation 3. Match rating update

$$\theta_i := \theta_i + K(R_{ij} - P(R_{ij} = 1))$$

The used probability function can be seen as a reparameterization of the Bradley-Terry model for pair-wise comparisons [24]. Under the Bradley-Terry model if two objects have true ratings $\pi 1$, $\pi 2$, then the first object is preferred (will rank higher in comparison) with probability $\pi 1/(\pi 1 + \pi 2)$. Instead of the logistic function, it is possible to use a normal cumulative distribution, which adheres to the Thurstone-Mosteller model for paired comparisons [39].

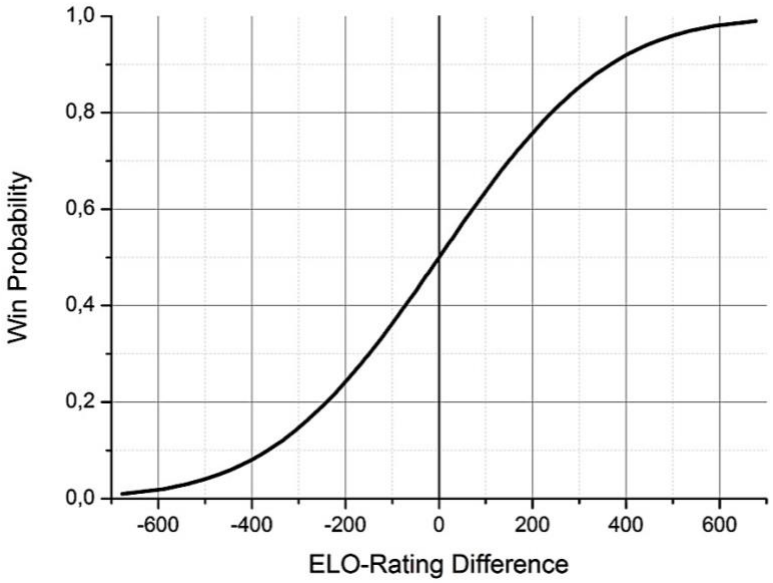



Figure 1. Winning expectancy curve

Since the logistic function and normal cumulative distribution function have almost identical shapes, the deviation between these two variants is in practice not essential. Contemporary realizations of the Elo rating system

generally use the logistic function because it is simpler to utilize in practical applications.

The value of the constant *K* in the update rule (see Equation 3) determines the behavior of the system. If *K* is small, the estimation converges too slowly, if K is large, the estimation is unstable as it gives too large a weighting to the last few attempts.

This study follows the official World Chess Federation ratings which use a tiered *K*-factor system, that denotes the players could have different K-factors:

- *K=40* for new players until they play 30 games
- *K=20* for players with > 30 games and never had an *ELO* > *2400*
- *K=10* for players with > 30 games and have had an *ELO* > *2400*

This system asserts a margin of uncertainty for the ratings of new players, facilitating them to reach their appropriate ELO in a short period. It also cushions extremely skilled players from losing ELO to unfortunate one-off losses.

This sort of implementation can be seen in competitive video games such as League of Legends or Overwatch, where the players must play at least ten placement games before their ELO becomes publicly available. Considering the ease of implementation as well as of all the facts as mentioned above, this study uses ELO and the uncertainty modifier to calculate the skill rating levels of each player.

### 3.1.2 Flow Short Scale

Csikszentmihalyi [50], [51] studied what makes experiences enjoyable to people. He was interested in people's inner states while pursuing challenging activities, yet appear to be intrinsically motivating, that is, contain rewards in themselves – chess, rock climbing, dance, sports. In later studies, he investigated ordinary people in their everyday lives, asking them to describe their experiences when they were living life at its fullest and were engaged in pleasurable activities. He discovered that central to all these experiences was a psychological state he called flow, an optimal state of enjoyment where people are completely absorbed in the activity. Flow is a state where someone's skills are well balanced with the challenges posed by a task. It is characterized by a deep concentration on the task at hand, a perceived sense of control over actions, a loss of preoccupation with self, and the transformation of one's sense of time. The figure below (see Figure 2) depicts the mental state in terms of challenge level and skill level, according to Csikszentmihalyi's flow model.

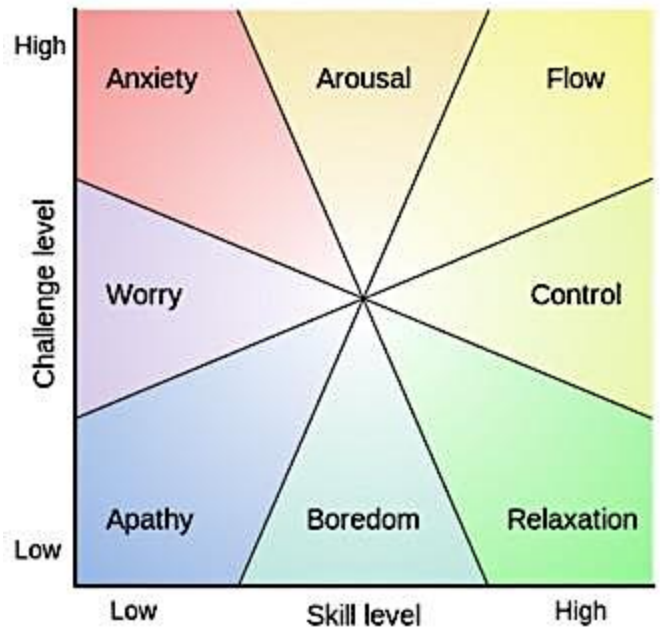



Figure 2. Csikszentmihalyi's flow model (Beatson, 2015)

Flow certainly sounds familiar to frequent players of computer games. Digital games provide players with an activity that is goal-directed, challenging, and requiring a set of skills. Most games offer immediate feedback on distance and progress towards the goals and objectives, though for instance, scorekeeping, status information (e.g., a health indicator), or direct in-game feedback. When a game is effective, the player's mind can enter an almost trance-like state in which the player is entirely focused on playing the game, and everything else seems to fade away - a loss of awareness of one's self, one's surroundings, and time. It is the experience that is strongly connected to what gamers and game reviewers commonly refer to as the "gameplay" of a game, i.e., the somewhat ambiguous term describing a holistic gaming experience, based on a fluent interaction with all active gaming elements, the progression of challenges offered, and the ability of a game to continuously command the attention of a player.

Sweetser and Wyeth [52] have adopted and extended Csikszentmihalyi's conceptualization of flow in their "Game Flow" model of player enjoyment, formulating a set of useful design criteria for achieving satisfaction in electronic games [53]. Csikszentmihalyi's original work on flow suggests that these peak experiences are quite rare – the exception rather than the rule. Nevertheless, the flow model of game enjoyment clearly illustrates the importance of providing an appropriate match between the challenges posed and the player's skill level. The flow experience can easily break down when the player's skills systematically outpace the challenges the game can offer (leading to boredom) or when game challenges become overwhelming in light of the available skills (resulting in frustration). Challenge is probably one of the most important aspects of good game design, and adjusting the challenge level to accommodate the broadest possible audience in terms of player motivation, experience and skill is a significant challenge for current game designers.

Being able to detect frustration and boredom is of importance as indicators of when a person is not experiencing flow, but also, and perhaps more interestingly, because successful games strike a balance between positive and negative emotions [54] is in line with the view that games are often being designed to develop a negative emotion in the face of challenge, only to be followed by a positive emotional peak when the challenge is overcome [55]. This idea leads to a richer, more exciting gaming experience and can be illustrated with a flow wave diagram below (see Figure 3). In sum, behavioral indicators of involvement or interest are required, as well as indicators of both boredom and frustration.

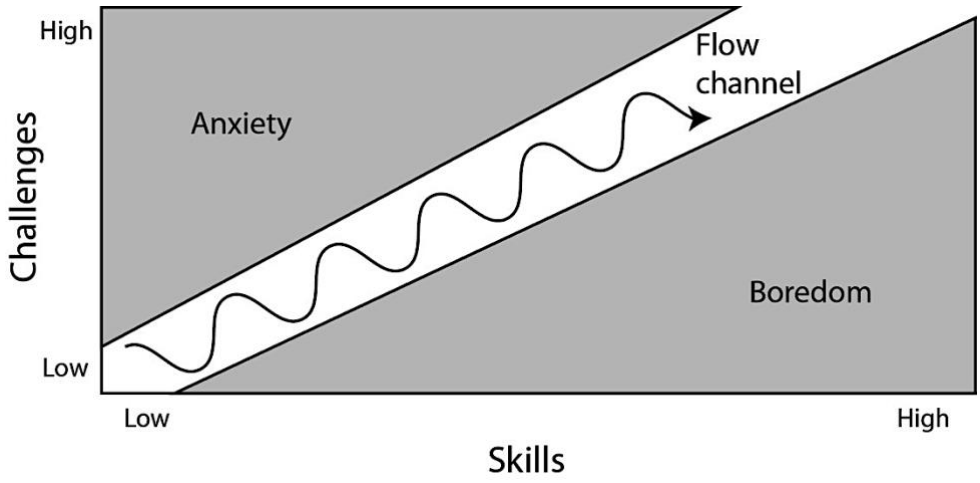



Figure 3. Game flow wave diagram (Csikszentmihalyi, 1990)

Due to the nature of this instrument's ability to assess the quintessential aspects of a game such as Anxiety and Challenge, this dissertation considered this as one of the study's primary instruments to determine the effects of "Flow" in a competitive environment where variance in skill levels are involved.

### 3.1.3 Self-Assessment Mannequin

The Self-Assessment Manikin (SAM) developed by Bradly and Lang [56] is a non-verbal pictorial assessment technique that directly measures the pleasure(valance), arousal, and dominance associated with a person's affective reaction to a wide variety of stimuli. Hence, this is an inexpensive and easy method for quickly assessing reports of affective response in various perspectives ranging from the circumplex model of affect [57] for identifying emotions experienced to evaluate positive and negative affective states of an individual [58].

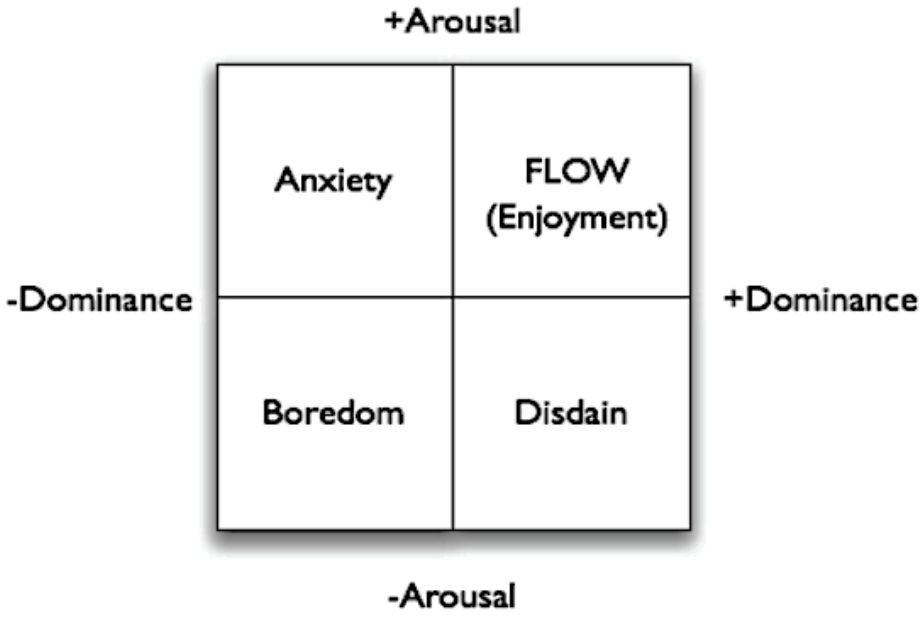



Figure 4. Affective mapping of flow channels (Gilroy et al. 2009)

In the context of Flow, Gilroy et al. [59], describe their framework that dispenses with Csikszentmihalyi's original model [60] mapping challenge and skill to Arousal and Dominance as depicted above in diagram (Figure 4). Moreover, these are the two of the three key outcomes of the Self-Assessment Manikin. Thus, this dissertation utilizes it to reinforce the results.

### 3.1.4 Positive and Negative Affect Scale

The Positive and Negative Affect Schedule (PANAS) is a 20 item self-reported measure of positive and negative affect developed by Watson, Clark, and Tellegen [61]. NA and PA reflect dispositional dimensions, with high-NA epitomized by subjective distress and unpleasurable engagement, and low NA by the absence of these feelings. By contrast, PA represents the extent to which an individual experiences pleasurable engagement with the environment. Thus, emotions such as enthusiasm and alertness are indicative of high PA, while lethargy and sadness characterize low PA [62]. It has, however, been argued that the labels, positive affect, and negative affect are misleading. Watson, Wiese, Vaidya, and Tellegen [63] point out that PA and NA are predominantly defined by the activation of positively and negatively valenced affects, respectively (its absence typifies, i.e., the lower ends of each dimension).

Thus, PA and NA can be paired with the results of the Self-Assessment Manikin (SAM), which has Valance as one of the three outcomes which directly correlates with Affect (PA with Positive Valance and NA with Negative Valance).

### 3.1.5 NASA Task Load Index

NASA Task Load Index (TLX). The NASA Task Load Index [64] uses six dimensions to assess mental workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. Twenty step bipolar scales are used to obtain ratings for these dimensions. A score from 0 to 100 (assigned to the nearest point 5) is obtained on each scale.

A weighting procedure is used to combine the six individual scale ratings into a global score; this procedure requires a paired comparison task to be performed before the workload assessments. Paired comparisons require the operator to choose which dimension is more relevant to workload across all pairs of the six dimensions. The number of times a dimension is chosen as more relevant is the weighting of that dimension scale for a given task for that operator.

The development of the TLX has implied an essential and vast program of laboratory research [65], and the instrument's sensitivity has been demonstrated using a great variety of tasks. TLX has been applied successfully in different multitask contexts, for example, in real [66] and simulated flight tasks [67]–[70]. Sawin and Scerbo [71] used the TLX technique to analyze the effects of instruction type and boredom proneness on vigilance task performance.

These characteristics of this instrument are widely used along with the Flow Short Scale to design video games [72] as well as to measure engagement in video games through cognitive and affective dimensions [73]. Hence, this work incorporates NASA TLX in order to measure the perceived workload in a refined manner.

### 3.1.6 BORG Rating of Perceived Exertion

The Borg category scale [74] is designed to describe perceptions of physical exertion during physical activities and is widely used to assess whole-body exertions. The scale consists of numbered categories, 6–20, and verbal anchors, from "very, very light" to "very, very hard" to increase the usability of the scale.

This scale has been extensively studied along with NASA TLX as an anchoring instrument [75] to distinguish the ambiguity between perceived mental workload and perceived physical workload.

### 3.1.7 Heart Rate

Heart Rate (HR) measurement in this study was conducted using a wearable photoplethysmography (PPG) sensor in the form of smartwatches to record and stream the HR data at 1Hz directly to the game as well maintaining a local log for redundancy. When compared with the gold-standard electrocardiography (ECG) sensors, it has been shown that PPG sensors possess a very high accuracy for measuring HR even in complex conditions such as exercising [76] as well as the situation where there are electrical interferences [77]. As seen below (Figure 5), during preliminary tests, the newer PPG sensors from Moto 360 performed relatively well in comparison to the older G Watch R, which dropped a significant amount of HR data.

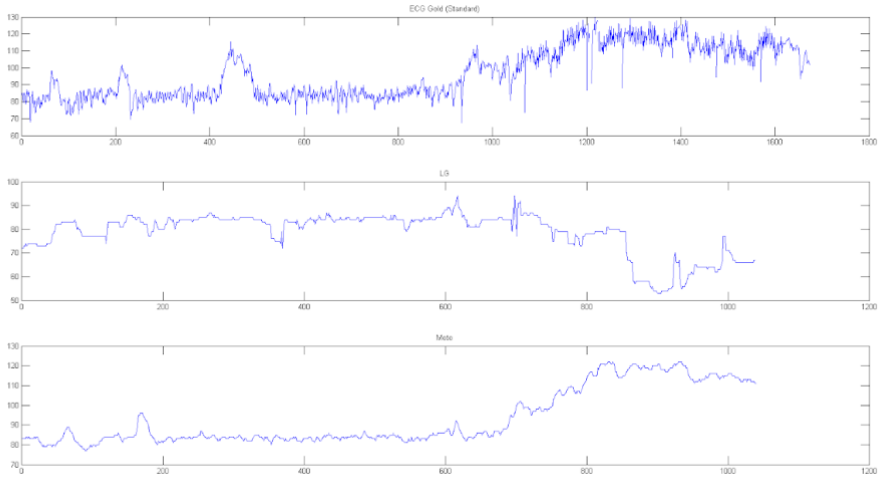



Figure 5. Comparison of ECG, G Watch R and Moto 360

Several studies on heart rate and perceived exertion ratings have concluded that ratings of perceived exertion can be used to gauge the physiological demands (HR in particular) of various physical activities [78]–[80]. Also, several other studies revealed that the same physiological demands have a strong correlation not only with the perceived physical exertion but also with perceived mental workload [81], [82], which can be tied to NASA TLX as well as the BORG rating of perceived exertion.

### 3.1.8 Movement

The player movement data acts as the non-conventional input method, and it is a crucial part of the game. In order to gather player movement data, this study employed the Kinect 2.0 sensor by Microsoft [83], which also tracked the positional data (depth in my case) of players' waist to be utilized in-game.

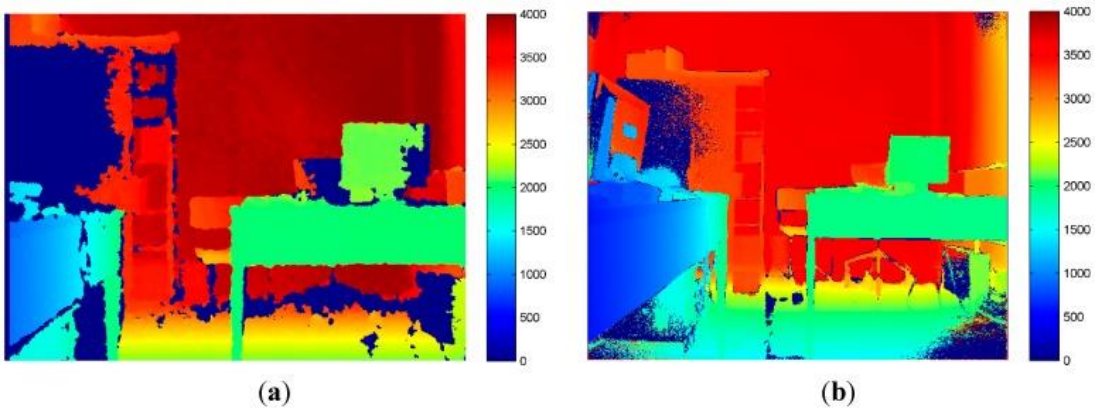



Figure 6. Depth maps by Kinect 1.0 (a) and Kinect 2.0 (b)

In the above image (Figure 6) In dark blue are represented the no-data value delivered by the sensors. The reason for choosing Kinect 2.0 over the Kinect 1.0 is due to the improved field of view (FOV) as well as depth mapping capabilities, which allows slightly better granularity in measuring depth [84].

## 3.2 ACTIVITY

For the activity, this study evaluated several competitive yet straightforward activities such as Chess, Tetris, Table Tennis, Pong, etc. The reasons behind these choices are that games, in general, promote competitiveness among players as well as require fewer resources compared to outdoor activities.

At first, the study was assessing games of chess. However, several difficulties were encountered while acquiring participants due to the time commitment towards each game. An average time per game was around 45 minutes. Hence, the study moved onto Tetris, where the game is simple yet does not have any complex mechanics that might be significant for the final results. Although this time, the game duration was relatively low, it was not appealing to the general diaspora of researchers who were willing to participate in the study. Once again, with the idea of Pong as well as the potential participants expected some physical stimulus rather than being sedentary throughout the game.

These predicaments lead to the exploration of Table Tennis. In this case, the consensus among participants seemed to be well received. However, measuring in-game metrics were not feasible with the technology at disposal.

After discussing with the members from the research group as well as with the dissertation supervisor, the study was adopted to be an exergame named AptoPong (adaptive Pong) based on the classic game of Pong that was made for a 24-hour hackathon [85].

The game consists of simple mechanics such as moving laterally, in which the player controls a paddle that would deflect the ball to the opponent's side. Since the game involves mild exertion as well as unfamiliar yet simple inputs, it suited the goal of the study, which is to measure the prolificacy of players using multiple instruments while avoiding previous experience related issues that the other games might have possessed.

## 3.3 TECHNOLOGY

### 3.3.1 Hardware Selection Criteria

The following are the proposed requirements to develop and perform the study:

**Computer**

- **Operating System:** 64-bit Windows 7, Windows 8.1, Windows 10.
- **Processor:** Intel Core i5-4430 or equivalent.
- **Memory:** 8 GB RAM.
- **Graphics:** NVIDIA GeForce GTX 960 2GB or equivalent.
- **Storage:** 800 MB of available disk space.

**Projector**

- **Resolution:** Full HD (1920 x 1080 px)
- **Aspect Ratio:** 16:9
- **Contrast Ratio:** 100,000:1
- **Image Format:** 60" - 100"
- **Interface:** HDMI or Display Port

For the hardware equipment, the study employs the PEPE platform from the Augmented Human Assistance project [86], which satisfied all of the abovementioned criteria as well as it was in the disposal at the NeuroRehabilitation Lab at the time of my study.

### 3.3.2 Multimodal Data Inquiry

#### Motion Sensing

- Microsoft Kinect 2.0

#### Smartwatch

- **Sensors:** PPG, Accelerometer, Gyroscope
- **Operating System:** Android Wear 2.0

### 3.3.3 Software

#### Game Engine

- Unity 3D

#### Machine Learning

- Unity ML-Agents Toolkit
- TensorFlow

#### Data Analysis

Jupyter (Python)

#### Python Packages

- Matplotlib
- Numpy
- Pandas
- Scikit-learn
- Scipy

## 3.4 DESIGN & DEVELOPMENT

### 3.4.1 The Game

As discussed in the Activity section (3.2) of this chapter, the game was initially conceived in a 24-hour hackathon [85]. This study has made several significant changes to the original game to facilitate the integration of data collection.

In the context of design, it was ideated to be an exergame to be played by two players or against an AI opponent (section 3.4.2). Furthermore, we decided to project the game screen to the floor and to use the player's lateral movement of the body as the only input in place of traditional input methods to promote exercise.

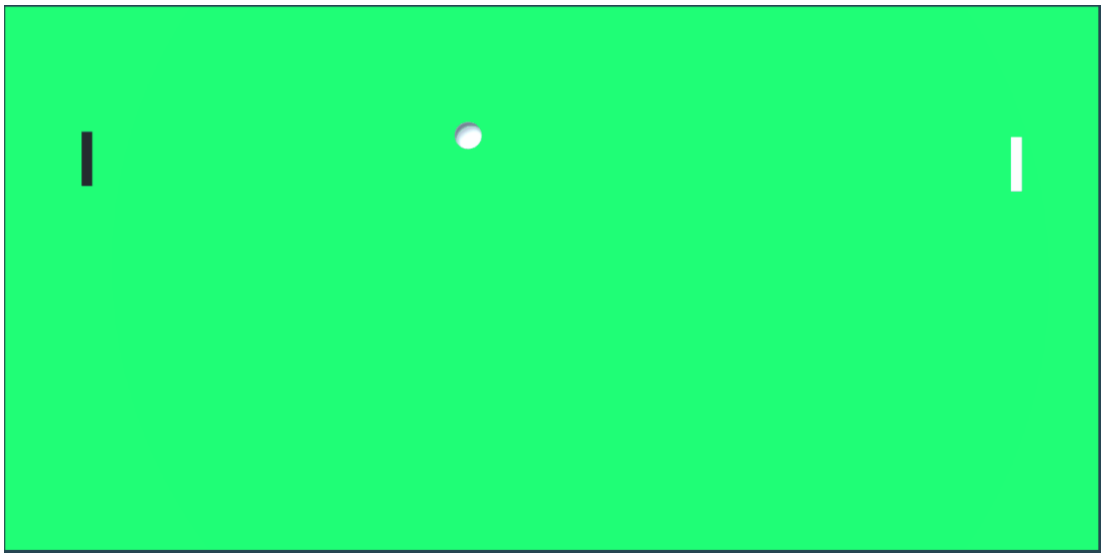


Figure 7. Top-down view of AptoPong

### 3.4.2 The AI

*State Machines*

In a simple game like ours, the AI implementation is usually done by utilizing the State Machines, which is the fundamental element of State design pattern in software engineering. There are two types of state machines: Finite State Machine (FSM) and Infinite State Machine. The FSM is composed of a finite number of states, transitions, and actions that can be modeled with flow graphs, where the path of logic can be detected when conditions are met [87].

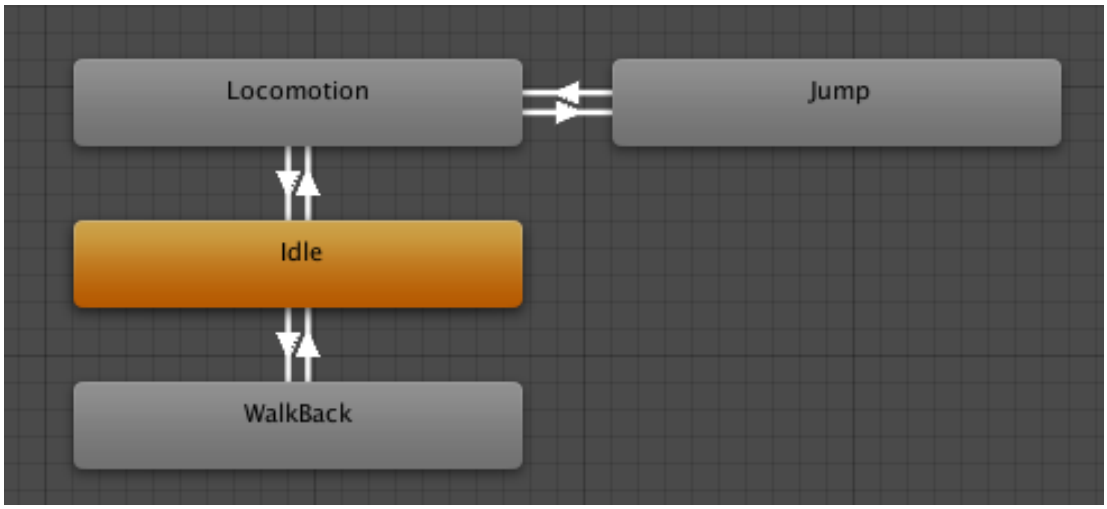



Figure 8. Simple state machine of movement

An FSM stores the status of something at a time and can only be in one of the finite number states at any given time. The status changes based on external inputs, as well as the shift from one state to another, is called a transition. A finite state machine is defined by a list of its states, its initial state, and the conditions for each transition.

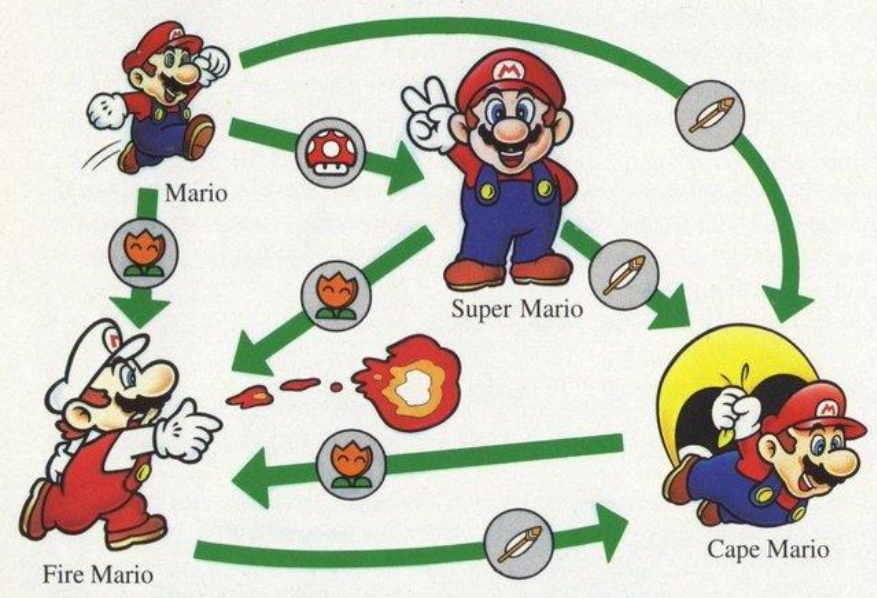



Figure 9. Mario's transitional states

As seen in Figure 8, this FSM depicts the simple movement states of a character within a game environment as well as the transitional paths it can take to go from one state to another. Another example of a familiar game scenario of Mario (Figure 9), which shows his various states and transitional paths. However, with an increasing number of states as wells as transitional paths, things can get overwhelming for a developer to program is one of the reasons that the traditional AI in videogames was unwieldy in comparison to the current video game AIs. Moreover, another reason to steer away from FSM for applying AI behavior in games is that the AI might be too unforgiving or too predictable due to the strict states defined by the developer.

*Machine Learning*

The developers of Unity Game Engine (UGE) have provided the Machine Learning Agents toolkit for facilitating the integration of ML aspects to the Unity Editor.

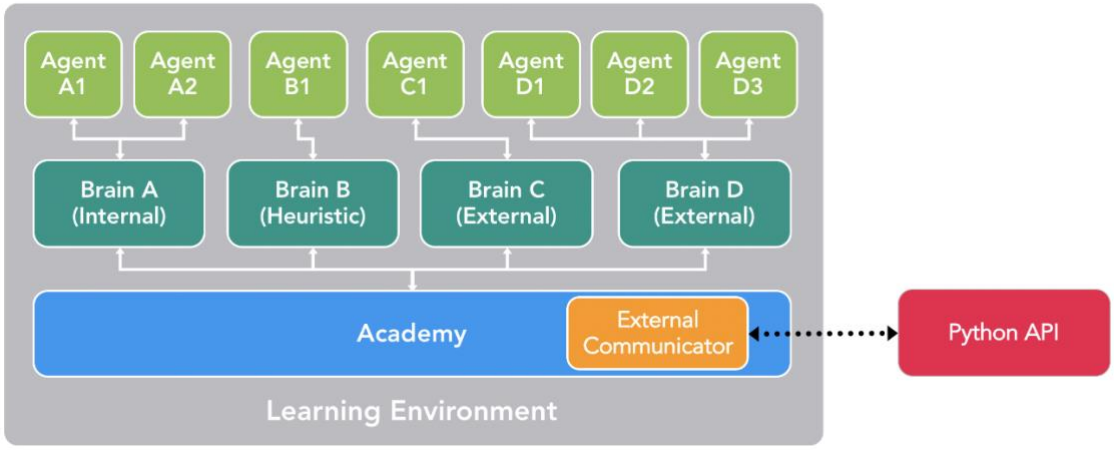



Figure 10. The learning environment of the Unity Editor and the Python interface.

The ML-Agents toolkit is an open-source project which enables researchers and developers to build simulation environments using the Unity Editor and interact with them by utilizing Python API [88]. As seen above, the toolkit consists of two components ML-Agents SDK which is imported into a project, and a Scene can be made into a Learning Environment and an interfacing Python package where both of them benefit from all the properties of UGE. ML-Agents takes advantage of a reinforcement learning (RL) technique, which works with a reward/punishment mechanism, called Proximal Policy Optimization (PPO). PPO is the preferred training method that Unity has developed, which uses a Neural Network (NN) and is implemented in TensorFlow [89], which runs in a separate Python process and communicates to Unity as shown above (Figure 10).

For this game, PPO was preferred, since it was an out of the box solution for UGE and it required a minimal amount of coding compared to the traditional FSM based solutions. Initially, the study was utilizing a single agent scenario (single scene) with vector observations (collisions) while taking continuous action (moving the paddle up or down) and dense rewards (+1 for scoring, 0.5 for bouncing the ball on the paddle and -1 for scored against).

However, this process seemed to be too lengthy for accomplishing a decent model that is adept at the game. Thus, as depicted in the diagram below (Figure 11), I decided to use a hybrid approach of increasing the number of agents to 14 (7 for player A and 7 for Player B) for the same brain and academy to accelerate the training process.

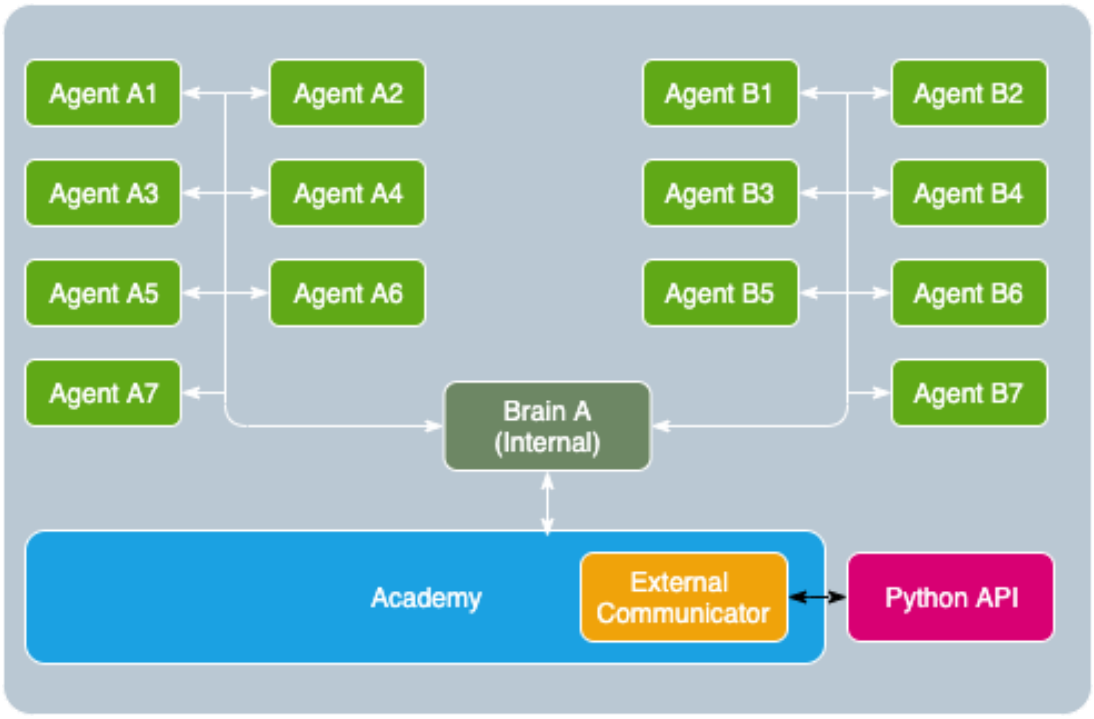



Figure 11. Optimized learning environment for my game

The new optimized approach made the learning considerably faster as well as keeping the process surprisingly stable. The image below (Figure 12) shows the training process with seven sets of agents playing against each other.

In order to perform this learning procees, the study utilized a gaming laptop with an Intel Core i7 7700HQ Processor, 32 GB of DDR4 2400MHz RAM, NVIDIA GeForce 1080 with 8GB VRAM and 512GB of SATA 3.0 SSD. After seven hours and 20 million iterations later, the model reached a stable state.

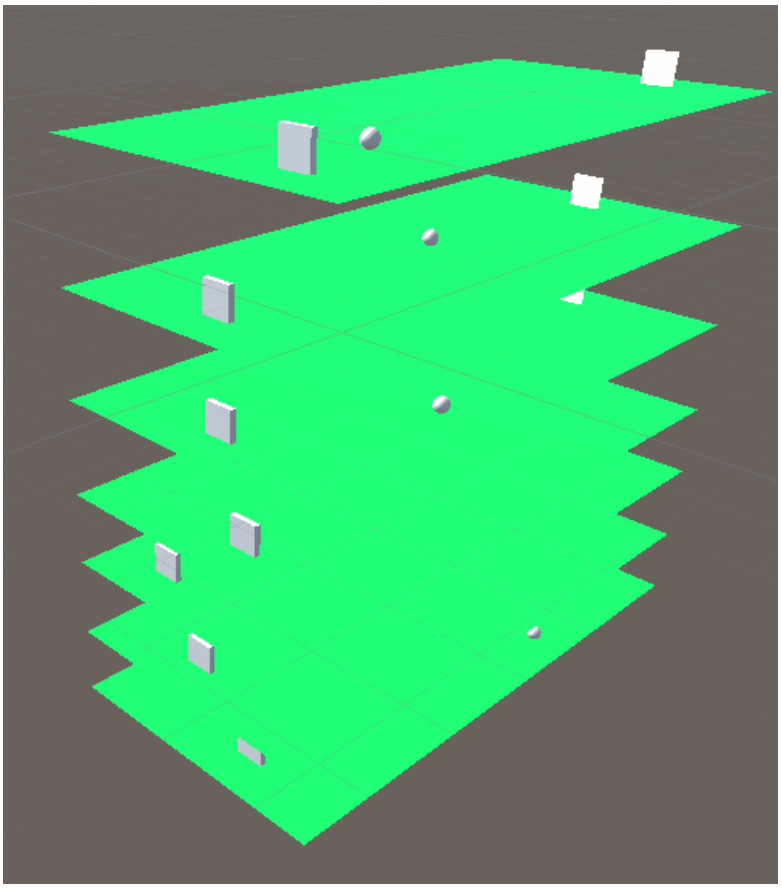


Figure 12. All 14 agents training in one environment

Furthermore, it was possible to confirm the stabilization of the model by analyzing the ML-Agent training toolkit statistics visualized in TensorBoard [89] using the works of Booth and Booth[90], Burda et al. [91], Juliani et al. [88] and Schulman et al. [92] as the guidelines. In below, the work presents the output of the TensorBoard along with the explanation for each graph based on the guidelines provided by the authors, as mentioned above.

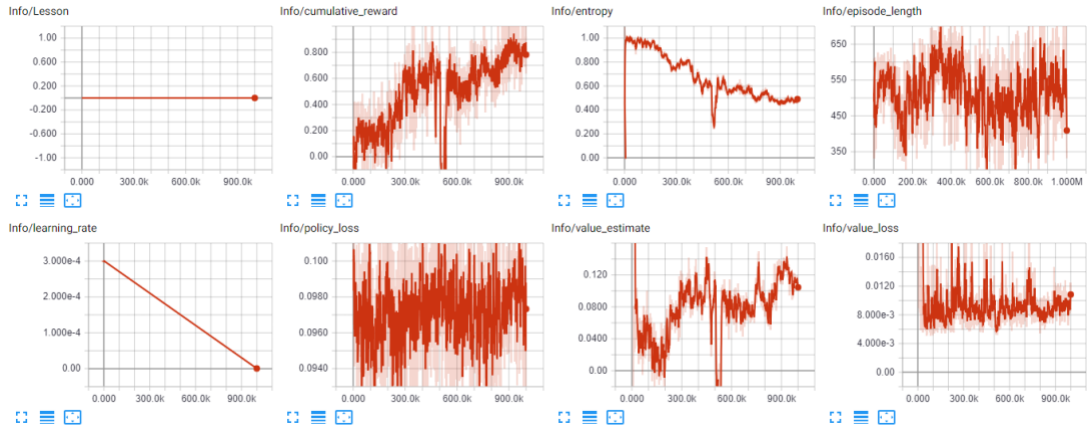



Figure 13. TensorBoard statistics of lesson, cumulative reward, learning rate, and policy loss

As illustrated in Figure 13, the TensorBoard output for Lesson (top left) must preferably plot the improvement from lesson to lesson. However, the significance of the graph only counts while performing curriculum training. The Cumulative Reward (top right) signifies the mean cumulative episode reward overall agents and ought to rise throughout a successful training session. The common tendency in reward should steadily grow over time, and inconsequential increase and decrease are to be expected. As expected from the suggestion of the Juliani et al. [88], depending on the complexity of the task, substantial growth in reward may not show up until millions of steps into the training process. The Learning Rate (bottom left) corresponds to how great the step the training algorithm takes as it finds for the most optimal policy, and it should superlatively decline overtime on a linear schedule. The Policy Loss (bottom right) implies the mean magnitude of the policy loss function, which correlates to how much the policy (course for deciding actions) is changing. Generally, these values will fluctuate during training and stay below 1.0.

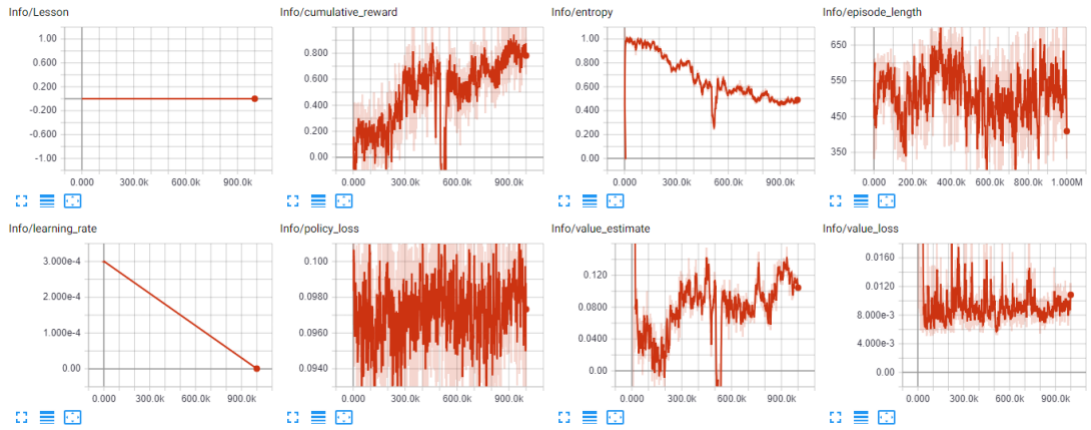



Figure 14. TensorBoard statistics of entropy, episode length, value estimate, and value loss

The Entropy (top left) denotes how arbitrary the decisions of a Brain are and should consistently decline during training and lowers the unpredictability of the model during the process. Episode Length (top right) represents the mean length of each episode in the environment for all agents. The Value Estimate (bottom left) implies the mean value estimate for all states call on by the agent and should increase as the cumulative reward increases and is related to the amount of future reward the agent predicts itself receiving at any given point. Value Loss (bottom right) represents the mean loss of the value function update and correlates to how well the model can predict the value of each state, and this ought to rise while the agent is learning, and then drops once the reward stabilizes.

As of these analyses, the model seemed to fit all the characteristics of a successfully trained and stabilized state.

### 3.4.3 Wearables

***PhysioSense***

The PhysioSense project is a part of the PhysioVR framework[93], which I helped in developing and co-authoring. It enables the framework to provide sensor data through User Datagram Protocol (UDP). PhysioSense consists of two applications, a mobile app for a smartphone, and a wearable application for Android wear compatible device, such as a smartwatch. Moreover, it accommodates all the available physiological and kinematic sensors to facilitate the PhysioVR framework.

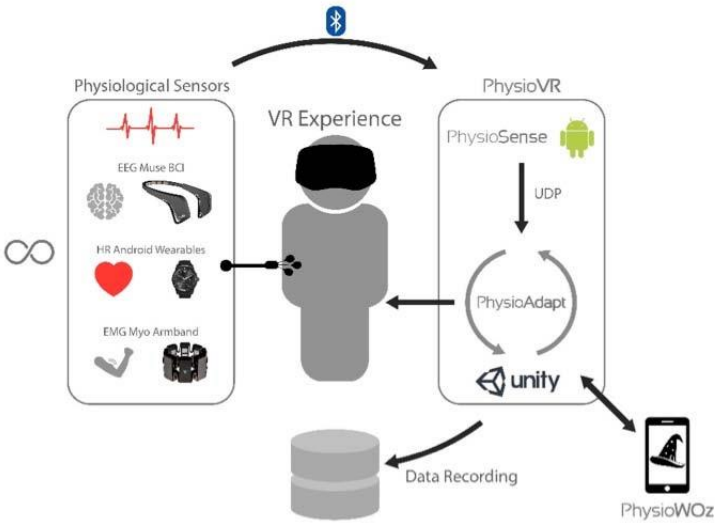



Figure 15. Conceptual architecture of PhysioVR and PhysioSense

This framework is compatible with any smartphone running Android Lollipop or later and Android Wear compatible device. Besides, the authors have also implemented and tested Muse: The Brain Sensing Headband[94] which is capable of providing raw frontal-lobe Electroencephalographic activity (EEG) metrics and Myo Armband: Gesture Control Armband which is capable of sensing musculoneural signals into machine-interpretable commands as well as providing raw Electromyographic activity (EMG).

***AptoSense***

AptoSense is a fork of PhysioSense which is optimized for longer battery life as well as low-latency scenarios. Also, AptoSense supports local logging as well as AndroidWear 2.0 based sensor suite such as motion, position, and environment-based sensors. However, this dissertation only utilizes HR, Accelerometer, and the Gyroscope, since other sensors fall beyond the scope of the study.

# Chapter 4: Methodology

This chapter describes the methods of the study. This chapter commences with the competition format (section 4.1.1). After that, section 4.1.2 refers to the usage of the questionnaires mentioned previously. Consequently, section 4.2 describes the choice of participants and the criteria of selection. Section 4.3 elaborates the reason for choosing a particular number of sessions, and finally, section 4.4 demonstrates the entire procedure of the study in detailed illustrations.

### 4.1.1 Competition Format

The overall format that is suitable for a competition depends on many factors, including the duration of the tournament (and the game), the mode of evaluation (e.g., win/loss/draw versus score-based), whether competitors have the chance to re-enter once they have lost and so on.

One of the simplest yet standard formats is single elimination (or knock-out), which consists of a succession of rounds where the winner of a single match progresses to the next round while the loser is eliminated. A slight modification of this is the double-elimination tournament, where a player has to lose two games to be disqualified. Another commonly used pairing system is the round-robin tournament: all competitors are paired against one another for one or more matches allowing each an equal opportunity to display their strength (although ordering might impact performances). However, the biggest issue with this approach is that it scales poorly, and a large number of competitors may take prohibitively long to evaluate. Since every competitor competes with every other opponent, the winner of a round-robin tournament is typically considered to depend much less on luck than of a single-elimination tournament.

Based on the abovementioned formats, for this study, the round-robin format was chosen not only due to the number of participants that were initially willing to participate but also the nature of the Elo rating system, which values every player facing each other at least once to generate a reliable rating. Moreover, in the case of the current study, the number of players is even ($n=10$), the well-known "circle design" performs well with respect to fairness [95].

### 4.1.2 Questionnaires

As of previously hypothesized research questions in section 1.3, this study employs questionnaires that could address the subjective experience of performing an activity (playing an exergame in my case) in a granular manner, such as providing insights to perceived workload, challenge, anxiety, exertion,

and flow, etc. The questionnaires to be used in the study are previously addressed in the Instruments section (3.1) of this dissertation.

## 4.2 PARTICIPANTS

For this study, 10 participants were invited based on the following requirements: Be between 15 to 80 years old. Present no conditions that may interfere with understanding, communication, and the execution of the task (exergame) as well as understand the English language and are motivated to participate. These requirements were mentioned in the consent form (see Appendix A: Consent Form), and the participants read and signed it after thoroughly understanding all the stated facts and procedures. Furthermore, all of the participants had previous experience with exergames.

The sample of participants for this study had the minimum age of 24, and the maximum age of 41, the median value of age was 29.0, and the standard deviation is 4.99 Out of the 10 participants, 8 of them were males, and 2 of them were females. Furthermore, the study also recorded the baseline HR of the players before every session.

## 4.3 SESSIONS

Although previous studies on Elo based rating system concluded that 25 games are needed to obtain a reliable Elo rating for a chess players[96], [97], in this particular case a round-robin style tournament with a short, yet fast-paced game, produced a clear separation among player skill ratings within as low as eight games with the adjusted uncertainty value of $K=40$ (refer Equation 3) for new players using simulations with the pre-trained model from my implementation of RL based on the PPO training. Thus, adhering to the round-robin schedule, each participant will be facing each opponent at least once with equally distributed intervals between each game.

## 4.4 PROCEDURE

First of all, the participants will go through the consent form (see Appendix A), clarify any doubts, and give their consent to participate in the study. This procedure is applicable for the first session only. From then on and in all other consecutive sessions, the participants will be wearing a composite sensor (smartwatch) as depicted in Figure 16, which measures heart rate using Photoplethysmography (PPG) sensors as well as spatial movement-related measurements using the built-in accelerometer and gyroscope on their non-dominant hand.

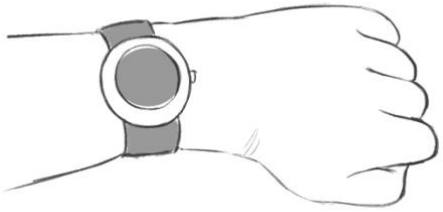


Figure 16. Smartwatch placement on the non-dominant hand

Subsequently, the participant's HR will be recorded for 1 minute to serve as their baseline HR for future data analysis. As in the case of the consent form for the first session, the participants will be going through an introduction and initial warm-up of the exergame (AptoPong) against the AI opponent trained with PPO, as seen below (Figure 17).

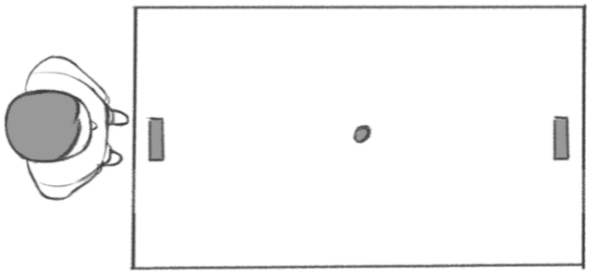


Figure 17. Participant against AI opponent

Afterward, the participant will be facing their appropriate opponent (human participants) from the round-robin tournament schedule, as shown below (Figure 18).

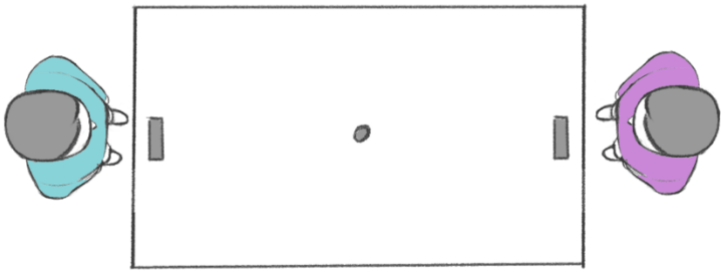


Figure 18. Initial position of the players

At the end of the first match, the participants will switch their starting sides, as depicted in the figure below (Figure 19), in order to eliminate any disadvantages of a dominant-side bias.

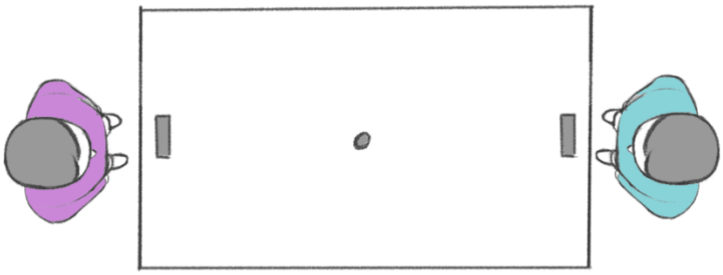


Figure 19. Switched starting position after the initial game

At the end of each session (facing an opponent), the participant will be requested to handover the smartwatch and will be given a battery of brief questionnaires (refer Appendix B, Appendix C, Appendix D, Appendix E, and Appendix F) to be answered before exiting the study area.

As shown below (Figure 20), the study setup consists of the floor projection, and the players are placed on the left and right. The hardware setup (PEPE) is placed on the top side of the projection, where the computer runs the game and projects through a short-throw projector attached to it.

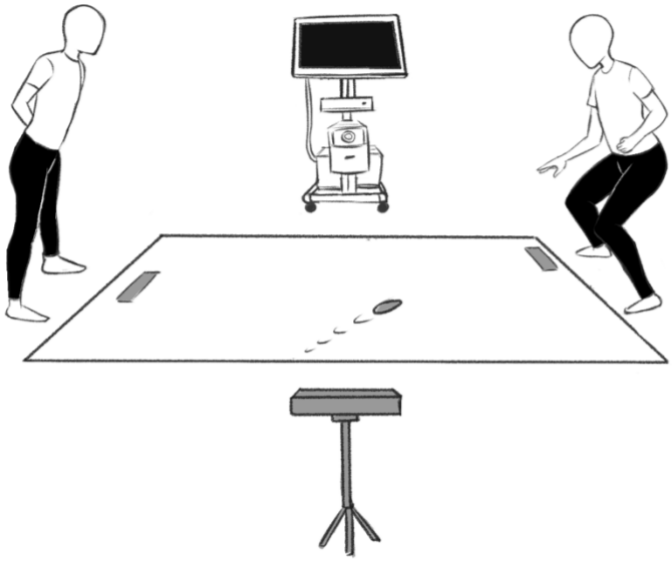


Figure 20. Study setup

The built-in Kinect sensor from PEPE was too close to the projection, and its field of view (of the IR sensor) could not accurately map the players on the 3D plane; hence I decided to add a Kinect 2.0 sensor at the bottom of the projection to negate this issue. As an advantage, this setup allows the study to be conducted even in confined spaces (see Figure 21).

Figure 21. Setup in confined space

# Chapter 5: Findings

This chapter commences with Section 5.1, which shows the detailed final scores of the tournament, and section 5.2 presents both game and session-based ELO ratings for each participant. Subsequently, section 5.3, highlights the correlations of the findings, and this chapter concludes with section 5.4, which consists of all further refining of the correlations using linear regression.

## 5.1 TOURNAMENT RESULTS

The results at the end of the round-robin tournament schedule are the following:

| Player | GP | Win | Draw | Lost | PF | PA | PD | PTS |
|---|---|---|---|---|---|---|---|---|
| **B** | 9 | 8 | 0 | 1 | 197 | 128 | 69 | 24 |
| **J** | 9 | 7 | 1 | 1 | 181 | 137 | 44 | 22 |
| **I** | 9 | 7 | 0 | 2 | 188 | 156 | 32 | 21 |
| **D** | 9 | 7 | 0 | 2 | 193 | 169 | 24 | 21 |
| **E** | 9 | 4 | 2 | 3 | 168 | 151 | 17 | 14 |
| **C** | 9 | 2 | 3 | 4 | 183 | 187 | -4 | 9 |
| **H** | 9 | 3 | 0 | 6 | 135 | 164 | -29 | 9 |
| **G** | 9 | 2 | 0 | 7 | 168 | 190 | -22 | 6 |
| **A** | 9 | 1 | 1 | 7 | 142 | 196 | -54 | 4 |
| **F** | 9 | 0 | 1 | 8 | 137 | 214 | -77 | 1 |

Table 2. Tournament results and standings

## 5.2 TOURNAMENT ELO

| Player | ELO (Per Game) | ELO (Per Session) |
|---|---|---|
| **B** | 1192 | 1107 |
| **J** | 1111 | 1092 |
| **I** | 1113 | 1071 |
| **D** | 1110 | 1074 |
| **E** | 1023 | 1016 |
| **C** | 980 | 967 |
| **H** | 967 | 961 |
| **G** | 936 | 923 |
| **A** | 847 | 895 |
| **F** | 879 | 864 |

Table 3. Tournament ELO rating

The table above (Table 3) displays the final ELO of the tournament, calculated in two different manners. Per Game-based ELO involves the calculation of ELO for each game. Per Session-based ELO is calculated based on the outcome of each session rather than individual games. Each session consists of at least two games and, on rare occasions, three, due to a mutually agreed tie-breaker or rematch.

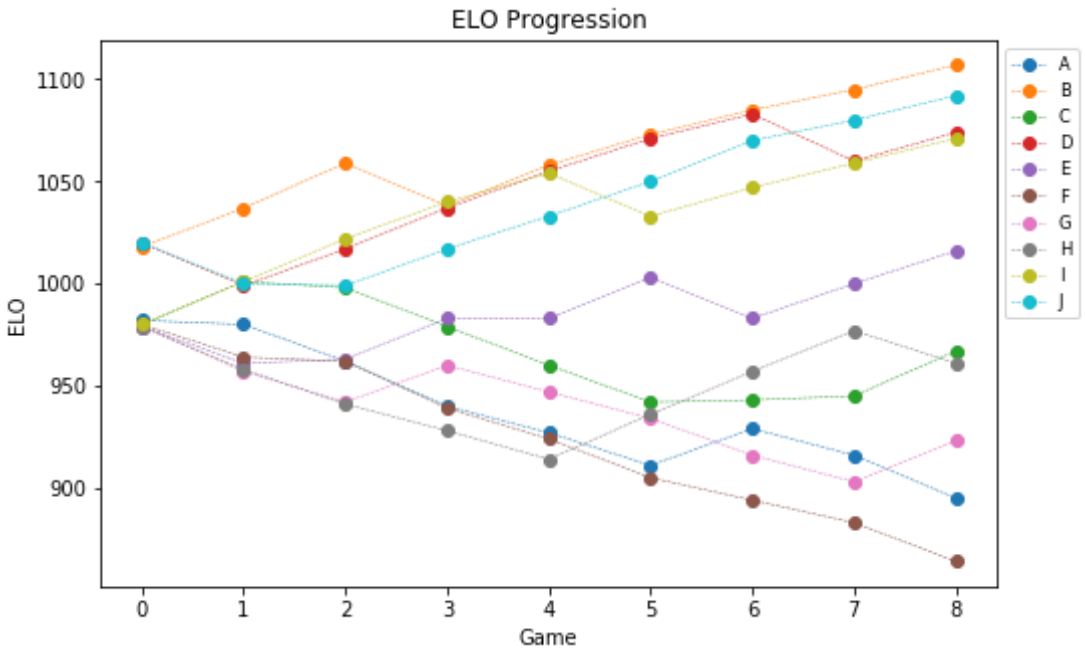



Figure 22. ELO Progression throughout the tournament

Furthermore, this dissertation was exploring the progression of ELO throughout the tournament to get an idea of how each player's skill level distinctively separates one from another. The figure above (Figure 22) shows the ELO progression throughout the tournament for each player. Please note that the starting ELO of 1000 is being ignored, and only the ELO outcomes of the matches are being plotted here.

Subsequently, in terms of a player's prolificacy, the study was surveying the distribution ELO ranges where each player falls into and how much variance does each player exhibited in terms of ELO. The figure below (Figure 23) depicts the clear separations between highly prolific players and the rest.

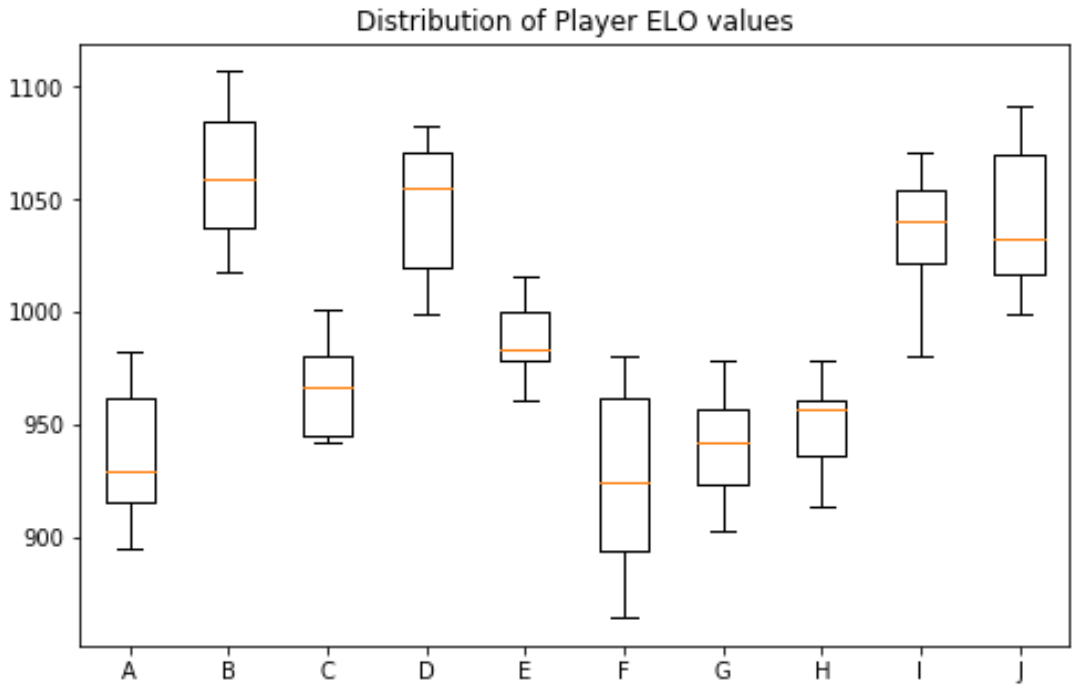



Figure 23. Distribution of Player ELO

Furthermore, the study was identifying for similar patterns in progression for other instrument variables; however, just by plotting them, no apparent patterns were noticeable among them. The plotted graphs can be found in Appendix I.

## 5.3 CORRELATIONS

To further analyze the dataset in hand, the study first tested the sample to see represents a normal distribution. The test was based on D'Agostino and Pearson's test [98], [99] that produce an omnibus test of normality. Once confirmed, the study proceeded to apply Pearson product-moment correlation coefficients for each player as well as all the instrument variables. The heatmap below (Figure 24) depicts the correlations in a comprehensible form. Individual heatmaps can be found in Appendix H.

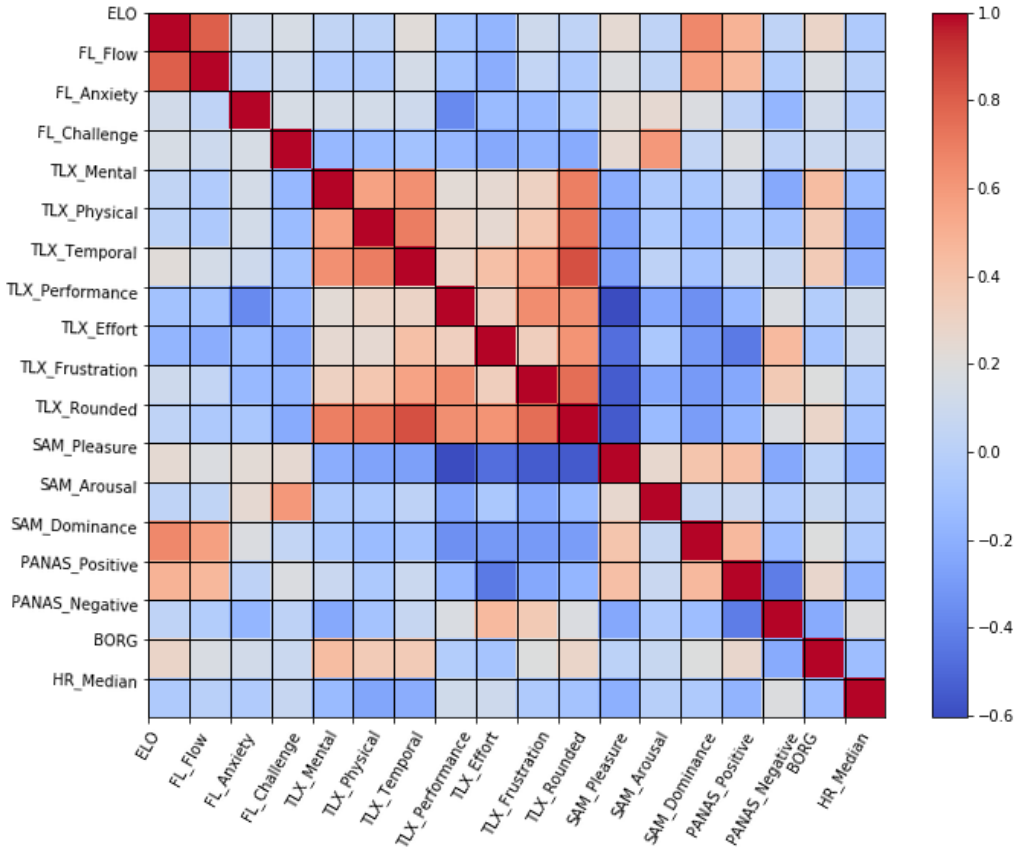



Figure 24. Correlation for all the players of all the variables

As discussed previously in this dissertation, ELO directly represents the prolificacy of a player in terms of their expertise/skill level. Hence, the study was focusing on the variables that correlate with ELO.

| Variable | ELO |
|---|---|
| **ELO** | 1 |
| **FL_Flow** | 0.802462 |
| **SAM_Dominance** | 0.664627 |
| **PANAS_Positive** | 0.490738 |
| **BORG** | 0.289993 |
| **SAM_Pleasure** | 0.242973 |
| **TLX_Temporal** | 0.220801 |
| **FL_Challenge** | 0.159397 |
| **FL_Anxiety** | 0.122413 |
| **TLX_Frustration** | 0.109351 |
| **TLX_Mental** | 0.042709 |
| **PANAS_Negative** | 0.030892 |
| **TLX_Rounded** | 0.030114 |
| **SAM_Arousal** | 0.029777 |
| **TLX_Physical** | 0.021563 |
| **HR_Median** | -0.04176 |
| **TLX_Performance** | -0.09782 |
| **TLX_Effort** | -0.17836 |

Table 4. Correlations with ELO

After examining the correlations shown above, ELO, Flow, and Dominance (SAM) were the three variables that had significantly high positive correlations among separate instruments. This correlation confirms that ELO (relative competitive skill) affects Flow and Dominance (subjective experiences) in a positive manner and addresses the ***RQ1. How relative competitive skill affects the subjective experience of players in gaming?***

Regarding ***RQ2. How subjective experience influences the absolute competitive skill of players in gaming?***, the absolute skill is assumed to be attributed to the tournament results (Table 2), and although the subjective experience correlated with these results, at certain skill level (refer players **I** and **D**), it seemed to be inconsistent.

## 5.4 REGRESSION

To further refine the outcomes from the correlations, this dissertation utilized linear regression to model the relationship between ELO and each variable obtained from the plethora of instruments that were used for this study. Figure 25 and Figure 26 graphically represent the outcome for the combinations of ELO and Flow.

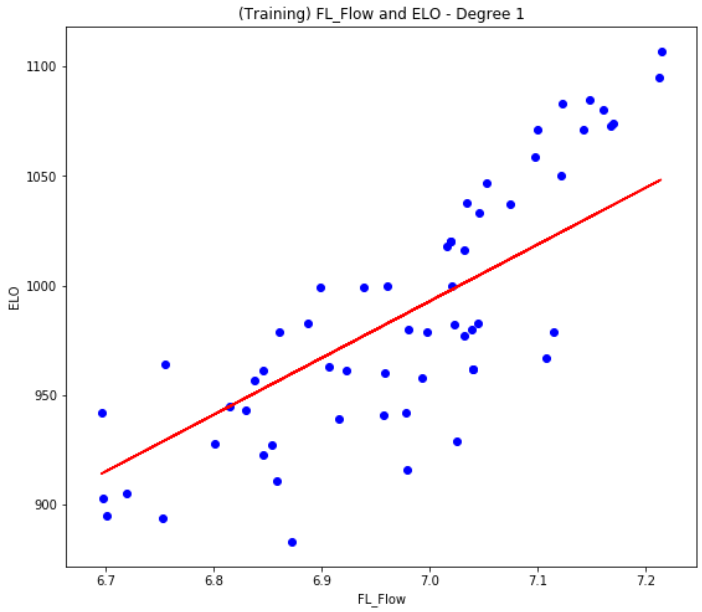



Figure 25. Training (Regression) for Flow and ELO

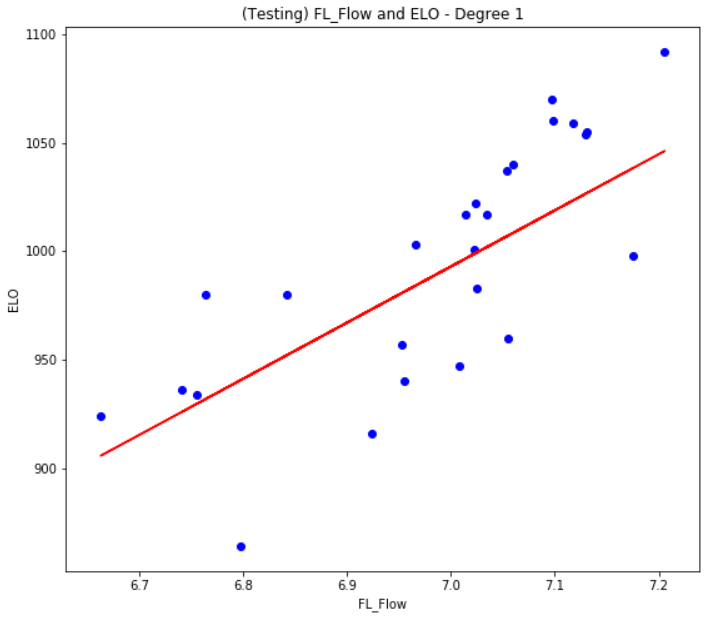



Figure 26. Testing (Regression) for Flow and ELO

Nevertheless, apart from FLOW, the rest of the variables were significantly less ideal for this study's objective, which is ***RQ3. Can the addition of subjective experience be beneficial for the traditional skill rating*** system? The results of training and testing are displayed on the table below (Table 5), and individual graphical representations can be found in Appendix J.

| | Training | Testing |
|---|---|---|
| **Flow** | 0.666109733 | 0.579084633 |
| **Anxiety (FLOW)** | 0.023974887 | -0.043126951 |
| **Challenge (FLOW)** | 0.049993896 | -0.090492842 |
| **Mental (TLX)** | 0.013574857 | -0.070139777 |
| **Physical (TLX)** | 0.000540195 | -0.039798608 |
| **Temporal (TLX)** | 0.030490232 | 0.038764084 |
| **Performance (TLX)** | 0.007864349 | -0.0273486 |
| **Effort (TLX)** | 0.062514261 | -0.08642549 |
| **Frustration (TLX)** | 0.000675366 | -0.027457559 |
| **TLX Rounded** | 9.75169E-07 | -0.040393613 |
| **Pleasure (SAM)** | 0.062093675 | 0.01460427 |
| **Arousal (SAM)** | 0.009083139 | -0.071111636 |
| **Dominance (SAM)** | 0.511720074 | 0.285820259 |
| **PA (PANAS)** | 0.252741011 | 0.192935818 |
| **NA (PANAS)** | 0.001320094 | -0.05376877 |
| **Exertion (BORG)** | 0.068357311 | 0.063105875 |
| **HR Median** | 0.000146104 | -0.037775296 |

Table 5. Regression Results

Thus, to solidify the previous outcome from the correlations, Flow is a quintessential component for predicting ELO.

## 5.5 MOVEMENT

Apart from the abovementioned findings, the study also revealed a pattern on the movement of the participants and their ELO ratings. Players who competed against closely skilled players tend to mirror their opponent's movement.


Figure 27. Movement data of the game between
Participant B (1107 ELO) and Participant J (1092 ELO)

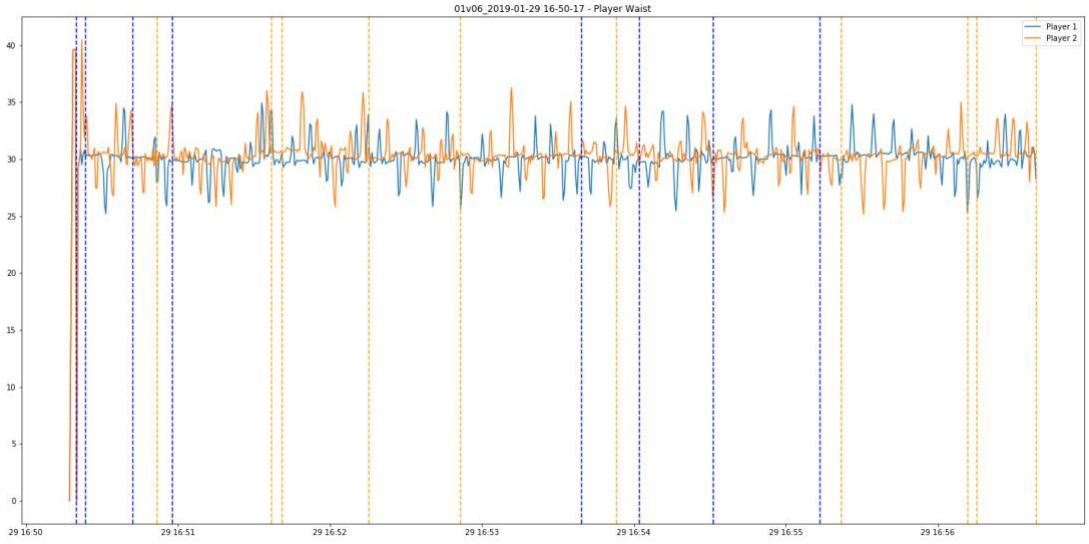



Figure 28. Movement data of the game between
Participant A (895 ELO) and Participant F (864 ELO)

As depicted above in Figure 27 (players with the highest ELO rating) and Figure 28 (players with the lowest ELO rating), regardless of the skill level of the players, closely matched opponents tend to mirror each other's movement and thus managing the pace and difficulty of the game in a mutual manner.

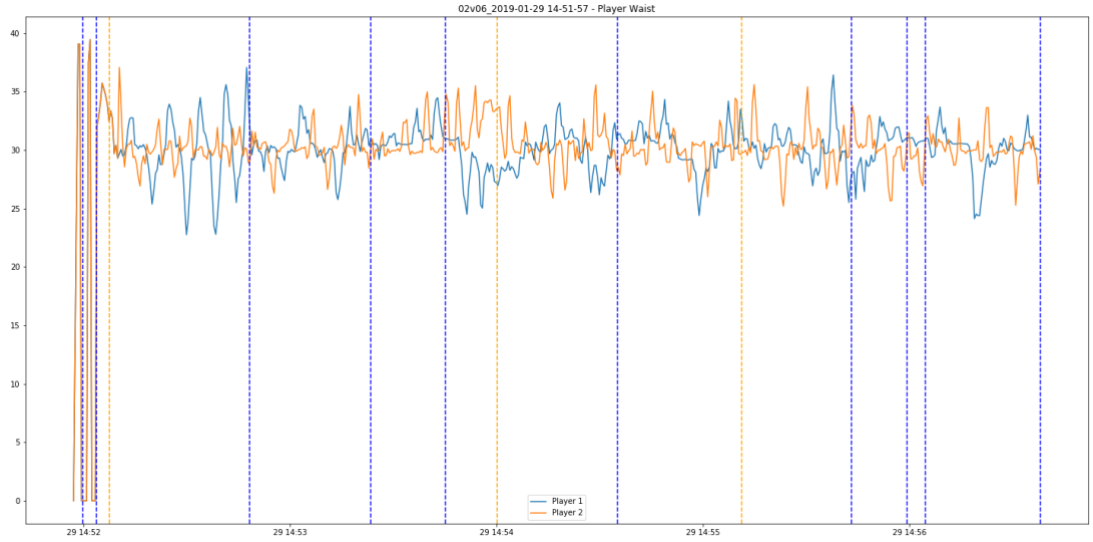



Figure 29. Movement data of the game between Participant B (1107 ELO) and Participant F (864)

However, when matching opponents who are on the opposite side of the rating spectrum, tend to move in a dissimilar manner. This phenomenon can be seen above in Figure 29, where the highest rated player and the lowest rated players were playing against each other.

# Chapter 6: Conclusions

This chapter begins with section 6.1, the observations and remarks based on the outcome of the study, and in section 6.2, The dissertation discusses the limitation of the study. The final section 6.3 presents the conclusion of this dissertation.

## 6.1 OBSERVATIONS AND REMARKS

The findings presented in this dissertation have demonstrated the correlation between prolificacy and several variables. Although these findings are not tested extensively, they offer some crucial implications for game user research and performance rating systems. For example, it would be valuable to determine what other psychophysiological factors might influence the expertise of an individual at a specific task. Based on the findings in this study, one might speculate that individuals who experience a relatively higher amount of flow might be able to perform better with a higher level of skill than their average peers.

Specifically, these individuals would demonstrate high levels of competence while engaged in an activity and be capable of working toward their objectives in that activity with greater ease and can be derived from highly skilled players' expressed mechanisms like automaticity and reduced cognitive load (TLX Mental) while being in Flow.

With respect to high prolificacy, researchers have explored that numerous characteristics differentiate highly skilled players from beginners, and they range from intuition, information processing speed to explicit prowess in a particular field [100]. These findings support the observation that the participants with considerably high Elo engage in exploring new ways to engage with the game (such as having a wider stance or controlling the paddle by merely tilting their hips) unlike less skilled participants (who were always doing what they learned at the beginning).

Besides, as mentioned previously, highly prolific participants reported higher levels of flow. However, it is unknown how players leverage the state of being in Flow in order to maximize their performance. Additional research is advocated to comment on this aspect of Flow in a competitive gaming scenario.

## 6.2 LIMITATIONS

As discussed above, the study managed to provide some insights into the research questions presented at the beginning of this dissertation. Nonetheless, regarding a few crucial concerns, it is suggested that these outcomes be interpreted with cautiousness. A fundamental issue is that most of the data that has been collected (except the in-game and physiological ones) in this dissertation were self-reported. While the use of self-report metrics is often pragmatic, there are known disputes with the accuracy and validity of these data. Notably, it is impossible to determine the extent to which participants performed and experienced, as they indicated in the study.

Future studies may overcome these limitations by investigating competitive gaming by means of objective manners (e.g., implementing better physiological sensors, adding more in-game metrics, etc.). Future investigations might also study levels of expertise with less ambiguity by studying participants who are known to be skilled at a competitive game and have had their ratings calculated by an accredited system (such as an official rating system of the game) to indicate prolificacy.

In addition to the nature of the data, there are other limitations to this investigation. The sample size of the study only consists of 10 participants. Although this number is enough for determining Elo ratings, this sampling neither adequately reflects players who might demonstrate significant variability in their skill level nor does it generalize the wide range of age groups. Because both access and time are consistent with increased levels of prolificacy, additional efforts to collect data from a larger pool of participants from different age groups for a more extended period of time is advisable.

## 6.3 CONCLUSION

Although there are many unanswered questions, this work considers these findings to be of significant value. Considering the role of psychophysiology in competitive gaming, each of these outcomes is arguably a highly desirable goal. For example, one would think it a great success if a player's psychophysiological metrics can predict the outcome of their future game, which directly signifies the skill level of the players.

Whereas this dissertation does not necessarily argue that these findings are generalizable to competitive gaming contexts, they may be leveraged in future competitive gaming applications. Research also indicates that prolificacy is related to the psychological state and development of performance through time by gaining experience [101].

Although traditional models of rating systems do not necessarily emphasize the importance of the psychophysiological state of the players, these findings indicate that it has a crucial influence on some cross-sections of the population.

Further, while this dissertation does not provide an alternative for the traditional rating systems, it shows the significance of considering other aspects of the competition, such as psychophysiological metrics to fine-tune the rating and potentially reveals more in-depth insight to the competition in comparison to just the binary outcome.

## 6.4 FUTURE WORK

This work can be further improved by tackling the issues mentioned in the limitations. One of the critical drawbacks was the lack of pre-session measurements from the participants. This issue can be addressed via applying pre-session questionnaires.

Furthermore, this would allow us to perform a causality analysis by examining both pre and post session measurements of the participants. Besides, as mentioned in the findings section, further investigation and training of physiological measures such as the movement of the players using Machine Learning techniques would vastly improve the accuracy of the rating system along with the use of ELO.

# Bibliography


[1] J. Hamari and M. Sjöblom, 'What is eSports and why do people watch it?', *Internet research*, vol. 27, no. 2, pp. 211–232, 2017.

[2] D. Lee and L. J. Schoenstedt, 'Comparison of eSports and traditional sports consumption motives.', *ICHPER-SD Journal Of Research*, vol. 6, no. 2, pp. 39–44, 2011.

[3] H. H. Bell and W. L. Waag, 'Evaluating the effectiveness of flight simulators for training combat skills: A review', *The international journal of aviation psychology*, vol. 8, no. 3, pp. 223–242, 1998.

[4] L. Panait, R. C. Merrell, A. Rafiq, S. J. Dudrick, and T. J. Broderick, 'Virtual reality laparoscopic skill assessment in microgravity', *Journal of Surgical Research*, vol. 136, no. 2, pp. 198–203, 2006.

[5] R. J. Scalese, V. T. Obeso, and S. B. Issenberg, 'Simulation technology for skills training and competency assessment in medical education', *Journal of general internal medicine*, vol. 23, no. 1, pp. 46–49, 2008.

[6] B. C. Witmer, J. H. Bailey, and B. W. Knerr, 'Training Dismounted Soldiers in Virtual Environments: Route Learning and Transfer.', ARMY RESEARCH INST FOR THE BEHAVIORAL AND SOCIAL SCIENCES ORLANDO FL ORLANDO …, 1995.

[7] J. E. Muñoz, T. Paulino, H. Vasanth, and K. Baras, 'PhysioVR: A novel mobile virtual reality framework for physiological computing', in *e-Health Networking, Applications and Services (Healthcom), 2016 IEEE 18th International Conference on*, 2016, pp. 1–6.

[8] J. E. Muñoz, E. Rubio, M. Cameirao, and S. Bermúdez, 'The Biocybernetic Loop Engine: an Integrated Tool for Creating Physiologically Adaptive Videogames', in *4th International Conference in Physiological Computing Systems. Presented at the PhyCS*, 2017.

[9] P. A. Alexander, C. T. Sperl, M. M. Buehl, H. Fives, and S. Chiu, 'Modeling domain learning: Profiles from the field of special education.', *Journal of educational psychology*, vol. 96, no. 3, p. 545, 2004.

[10] K. A. Ericsson and N. Charness, 'Expert performance: Its structure and acquisition.', *American psychologist*, vol. 49, no. 8, p. 725, 1994.

[11] S. Barab, M. Thomas, T. Dodge, R. Carteaux, and H. Tuzun, 'Making learning fun: Quest Atlantis, a game without guns', *Educational technology research and development*, vol. 53, no. 1, pp. 86–107, 2005.

[12] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer, 'The role of deliberate practice in the acquisition of expert performance.', *Psychological review*, vol. 100, no. 3, p. 363, 1993.

[13] P. K. Murphy and P. A. Alexander, 'What counts? The predictive powers of subject-matter knowledge, strategic processing, and interest in domain-specific performance', *The Journal of Experimental Education*, vol. 70, no. 3, pp. 197–214, 2002.

[14] R. S. Grabinger and J. C. Dunlap, 'Rich environments for active learning: A definition', *ALT-J*, vol. 3, no. 2, pp. 5–34, 1995.

[15] W. Peng, J.-H. Lin, and J. Crouse, 'Is playing exergames really exercising? A meta-analysis of energy expenditure in active video games', *Cyberpsychology, Behavior, and Social Networking*, vol. 14, no. 11, pp. 681–688, 2011.

[16] K. Harkness, *Official chess handbook*. D. McKay Co., 1956.
[17] I. Makarov, D. Savostyanov, B. Litvyakov, and D. I. Ignatov, 'Predicting Winning Team and Probabilistic Ratings in "Dota 2" and "Counter-Strike: Global Offensive" Video Games', in *International Conference on Analysis of Images, Social Networks and Texts*, 2017, pp. 183–196.
[18] N. Prakannoppakun and S. Sinthupinyo, 'Skill rating method in multiplayer online battle arena', in *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2016, pp. 1–6.
[19] G. G. Hall, J. M. Decelle, and L. R. O'Donnell, 'The Importance of Matchmaking in League of Legends and its Effects on Users', 2015.
[20] M. Claypool, J. Decelle, G. Hall, and L. O'Donnell, 'Surrender at 20? Matchmaking in league of legends', in *Games Entertainment Media Conference (GEM), 2015 IEEE*, 2015, pp. 1–4.
[21] M. E. Glickman, 'Parameter Estimation in Large Dynamic Paired Comparison Experiments', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, no. 3, pp. 377–394, Aug. 1999.
[22] A. E. Elo, *The rating of chessplayers, past and present*. New York: Arco Pub, 1978.
[23] E. Zermelo, 'Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung', *Mathematische Zeitschrift*, vol. 29, no. 1, pp. 436–460, 1929.
[24] R. A. Bradley and M. E. Terry, 'Rank analysis of incomplete block designs: I. The method of paired comparisons', *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.
[25] R. D. Luce, *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.
[26] R. Coulom, 'Computing "elo ratings" of move patterns in the game of go', *Icga Journal*, vol. 30, no. 4, pp. 198–208, 2007.
[27] S. Hacker and L. Von Ahn, 'Matchin: eliciting user preferences with an online game', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 1207–1216.
[28] W. Pieters, S. H. van der Ven, and C. W. Probst, 'A move in the security measurement stalemate: Elo-style ratings to quantify vulnerability', in *Proceedings of the 2012 New Security Paradigms Workshop*, 2012, pp. 1–14.
[29] A. Das Sarma, A. Das Sarma, S. Gollapudi, and R. Panigrahy, 'Ranking mechanisms in twitter-like forums', in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 21–30.
[30] C. S. Tsang, H. Y. Ngan, and G. K. Pang, 'Fabric inspection based on the Elo rating method', *Pattern Recognition*, vol. 51, pp. 378–394, 2016.
[31] D. A. Reid and M. S. Nixon, 'Using comparative human descriptions for soft biometrics', in *2011 International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–6.
[32] D. A. Reid, M. S. Nixon, and S. V. Stevenage, 'Soft biometrics; human identification using comparative descriptions', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1216–1228, 2013.
[33] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter, 'Soft biometric recognition from comparative crowdsourced annotations', 2015.
[34] N. Almudhahka, M. Nixon, and J. Hare, 'Human face identification via comparative soft biometrics', in *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2016, pp. 1–6.

[35] N. Y. Almudhahka, M. S. Nixon, and J. S. Hare, ‘Unconstrained human identification using comparative facial soft biometrics’, in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016, pp. 1–6.
[36] P. C. Albers and H. de Vries, ‘Elo-rating as a tool in the sequential estimation of dominance strengths’, *Animal Behaviour*, pp. 489–495, 2001.
[37] E. T. Johnson, N. Snyder-Mackler, J. C. Beehner, and T. J. Bergman, ‘Kinship and dominance rank influence the strength of social bonds in female geladas (Theropithecus gelada)’, *International Journal of Primatology*, vol. 35, no. 1, pp. 288–304, 2014.
[38] N. Snyder-Mackler, J. N. Kohn, L. B. Barreiro, Z. P. Johnson, M. E. Wilson, and J. Tung, ‘Social status drives social relationships in groups of unrelated female rhesus macaques’, *Animal behaviour*, vol. 111, pp. 307–317, 2016.
[39] M. E. Glickman, ‘The glicko system’, *Boston University*, 1995.
[40] R. C. Weng and C.-J. Lin, ‘A bayesian approximation method for online ranking’, *Journal of Machine Learning Research*, vol. 12, no. Jan, pp. 267–300, 2011.
[41] M. E. Glickman, ‘Example of the Glicko-2 system’, *Boston University*, 2012.
[42] R. Edwards, *Edo historical chess ratings*. 2004.
[43] P. Dangauthier, R. Herbrich, T. Minka, and T. Graepel, ‘TrueSkill Through Time: Revisiting the History of Chess’, in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, USA, 2007, pp. 337–344.
[44] T. Graepel and R. Herbrich, ‘Ranking and matchmaking: Grouping online players for competitive gaming’, *Game Developer Magazine*, vol. 13, no. 9, pp. 25–34, 2006.
[45] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, ‘Factor graphs and the sum-product algorithm’, *IEEE Transactions on information theory*, vol. 47, no. 2, pp. 498–519, 2001.
[46] T. P. Minka, ‘A family of algorithms for approximate Bayesian inference’, Massachusetts Institute of Technology, 2001.
[47] X. Zhang, T. Graepel, and R. Herbrich, ‘Bayesian online learning for multi-label and multi-variate performance measures’, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 956–963.
[48] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich, ‘Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft’s bing search engine’, 2010.
[49] B. Schölkopf, J. Platt, and T. Hofmann, ‘TrueSkill™: A Bayesian Skill Rating System’, 2007.
[50] M. Csikszentmihalyi, ‘Flow. The Psychology of Optimal Experience. New York (HarperPerennial) 1990.’, 1990.
[51] M. Csikszentmihalyi and I. Csikszentmihalyi, *Beyond boredom and anxiety*, vol. 721. Jossey-Bass San Francisco, 1975.
[52] P. Sweetser and P. Wyeth, ‘GameFlow: a model for evaluating player enjoyment in games’, *Computers in Entertainment (CIE)*, vol. 3, no. 3, pp. 3–3, 2005.
[53] R. Holt and J. Mitterer, ‘Examining video game immersion as a flow state’, *108th Annual Psychological Association, Washington, DC*, 2000.

[54] N. Ravaja, M. Salminen, J. Holopainen, T. Saari, J. Laarni, and A. Järvinen, 'Emotional response patterns and sense of presence during video games: Potential criterion variables for game design', in *Proceedings of the third Nordic conference on Human-computer interaction*, 2004, pp. 339–347.
[55] K. Keeker, R. Pagulayan, J. Sykes, and N. Lazzaro, 'The untapped world of video games', in *CHI'04 Extended Abstracts on Human Factors in Computing Systems*, 2004, pp. 1610–1611.
[56] M. M. Bradley and P. J. Lang, 'Measuring emotion: the self-assessment manikin and the semantic differential', *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.
[57] J. A. Russell, 'A circumplex model of affect.', *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
[58] A. Mehrabian and J. A. Russell, *An approach to environmental psychology.* 1974.
[59] S. W. Gilroy, M. Cavazza, and M. Benayoun, 'Using affective trajectories to describe states of flow in interactive art', in *Proceedings of the International Conference on Advances in Computer Enterntainment Technology*, 2009, pp. 165–172.
[60] M. Csikszentmihalyi, *Creativity: flow and the psychology of discovery and invention*, 1st ed. New York: HarperCollinsPublishers, 1996.
[61] D. Watson, L. A. Clark, and A. Tellegen, 'Development and validation of brief measures of positive and negative affect: the PANAS scales.', *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.
[62] D. Watson and L. A. Clark, 'Negative affectivity: the disposition to experience aversive emotional states.', *Psychological bulletin*, vol. 96, no. 3, p. 465, 1984.
[63] D. Watson, D. Wiese, J. Vaidya, and A. Tellegen, 'The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence.', *Journal of personality and social psychology*, vol. 76, no. 5, p. 820, 1999.
[64] S. G. Hart and L. E. Staveland, 'Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research', in *Advances in Psychology*, vol. 52, Elsevier, 1988, pp. 139–183.
[65] S. G. Hart and L. E. Staveland, 'Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research', in *Advances in psychology*, vol. 52, Elsevier, 1988, pp. 139–183.
[66] R. J. Shively, M. R. Bortolussi, V. Battiste, S. G. Hart, D. D. Pepitone, and J. H. Matsumoto, 'Inflight evaluation of pilot workload measures for rotorcraft research', 1987.
[67] V. Battiste and M. Bortolussi, 'Transport pilot workload: A comparison of two subjective techniques', in *Proceedings of the Human Factors Society Annual Meeting*, 1988, vol. 32, pp. 150–154.
[68] M. Nataupsky and T. S. Abbott, 'Comparison of workload measures on computer—generated primary flight displays', in *Proceedings of the Human Factors Society Annual Meeting*, 1987, vol. 31, pp. 548–552.
[69] P. S. Tsang, 'Cognit ve Demands in Automcition', *Aviation, space, and environmental medicine*, vol. 60, pp. 130–5, 1989.
[70] M. A. Vidulich and M. R. Bortolussi, 'A dissociation of objective and subjective workload measures in assessing the impact of speech controls in advanced helicopters', in *Proceedings of the Human Factors Society Annual Meeting*, 1988, vol. 32, pp. 1471–1475.

[71] D. A. Sawin and M. W. Scerbo, ‘Effects of instruction type and boredom proneness in vigilance: Implications for boredom and workload’, *Human factors*, vol. 37, no. 4, pp. 752–765, 1995.
[72] D. Sharek and E. Wiebe, ‘Using flow theory to design video games as experimental stimuli’, in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2011, vol. 55, pp. 1520–1524.
[73] D. Sharek and E. Wiebe, ‘Measuring video game engagement through the cognitive and affective dimensions’, *Simulation & Gaming*, vol. 45, no. 4–5, pp. 569–592, 2014.
[74] G. A. Borg, ‘Psychophysical bases of perceived exertion’, *Med sci sports exerc*, vol. 14, no. 5, pp. 377–381, 1982.
[75] A. DiDomenico and M. A. Nussbaum, ‘Interactive effects of physical and mental workload on subjective workload assessment’, *International journal of industrial ergonomics*, vol. 38, no. 11–12, pp. 977–983, 2008.
[76] M. Prospero, ‘Who has the most accurate heart rate monitor’, *Tom’s Guide*, 2016.
[77] G. Lu, F. Yang, J. A. Taylor, and J. F. Stein, ‘A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects’, *Journal of medical engineering & technology*, vol. 33, no. 8, pp. 634–641, 2009.
[78] W. H. Sinclair, R. M. Kerr, W. L. Spinks, and A. S. Leicht, ‘Blood lactate, heart rate and rating of perceived exertion responses of elite surf lifesavers to high-performance competition’, *Journal of Science and Medicine in Sport*, vol. 12, no. 1, pp. 101–106, 2009.
[79] J. Scherr, B. Wolfarth, J. W. Christle, A. Pressler, S. Wagenpfeil, and M. Halle, ‘Associations between Borg’s rating of perceived exertion and physiological measures of exercise intensity’, *European journal of applied physiology*, vol. 113, no. 1, pp. 147–155, 2013.
[80] C. I. Abrantes, M. I. Nunes, V. M. MaÇãs, N. M. Leite, and J. E. Sampaio, ‘Effects of the number of players and game type constraints on heart rate, rating of perceived exertion, and technical actions of small-sided soccer games’, *The Journal of Strength & Conditioning Research*, vol. 26, no. 4, pp. 976–981, 2012.
[81] S. Miyake, ‘Multivariate workload evaluation combining physiological and subjective measures’, *International journal of psychophysiology*, vol. 40, no. 3, pp. 233–238, 2001.
[82] Y.-H. Lee and B.-S. Liu, ‘Inflight workload assessment: Comparison of subjective and physiological measurements’, *Aviation, space, and environmental medicine*, vol. 74, no. 10, pp. 1078–1084, 2003.
[83] Z. Zhang, ‘Microsoft kinect sensor and its effect’, *IEEE multimedia*, vol. 19, no. 2, pp. 4–10, 2012.
[84] D. Pagliari and L. Pinto, ‘Calibration of kinect for xbox one and comparison between the two generations of microsoft sensors’, *Sensors*, vol. 15, no. 11, pp. 27569–27589, 2015.
[85] R. Madeira Rádio e Televisão de Portugal-RTP, ‘Maratona de programação ACIN Hackathon “caça” talentos na Madeira’, *@rtppt*. [Online]. Available: https://www.rtp.pt/madeira/eventos/maratona-de-programacao-acin-hackathon-caca-talentos-na-madeira-_20326. [Accessed: 25-Aug-2019].
[86] H. Simão and A. Bernardino, ‘User Centered Design of an Augmented Reality Gaming Platform for Active Aging in Elderly Institutions.’

[87] E. Gamma, *Design patterns: elements of reusable object-oriented software*. Pearson Education India, 1995.
[88] A. Juliani *et al.*, 'Unity: A general platform for intelligent agents', *arXiv preprint arXiv:1809.02627*, 2018.
[89] M. Abadi *et al.*, 'Tensorflow: A system for large-scale machine learning', in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.
[90] J. Booth and J. Booth, 'Marathon environments: Multi-agent continuous control benchmarks in a modern video game engine', *arXiv preprint arXiv:1902.09097*, 2019.
[91] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, 'Large-scale study of curiosity-driven learning', *arXiv preprint arXiv:1808.04355*, 2018.
[92] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, 'Proximal policy optimization algorithms', *arXiv preprint arXiv:1707.06347*, 2017.
[93] J. E. Muñoz, T. Paulino, H. Vasanth, and K. Baras, 'PhysioVR: A novel mobile virtual reality framework for physiological computing', in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2016, pp. 1–6.
[94] C. A. Aimone, A. S. Garten, S. E. Grant, O. Mayrand, and T. Zimmermann, *Brain sensing headband*. Google Patents, 2016.
[95] W. Suksompong, 'Scheduling asynchronous round-robin tournaments', *Operations Research Letters*, vol. 44, no. 1, pp. 96–100, 2016.
[96] H. L. Van Der Maas and E.-J. Wagenmakers, 'A psychometric analysis of chess expertise', *American Journal of Psychology*, vol. 118, no. 1, pp. 29–60, 2005.
[97] M. Antal, 'On the use of elo rating for adaptive assessment', *Studia Universitatis Babes-Bolyai, Informatica*, vol. 58, no. 1, pp. 29–41, 2013.
[98] R. B. d'Agostino, 'An omnibus test of normality for moderate and large size samples', *Biometrika*, vol. 58, no. 2, pp. 341–348, 1971.
[99] R. D'AGOSTINO and E. S. Pearson, 'Tests for departure from normality. Empirical results for the distributions of b 2 and√ b', *Biometrika*, vol. 60, no. 3, pp. 613–622, 1973.
[100] M. T. Chi, R. Glaser, and M. J. Farr, *The nature of expertise*. Psychology Press, 2014.
[101] A. Lesgold, H. Rubinson, P. Feltovich, R. Glaser, D. Klopfer, and Y. Wang, 'Expertise in a complex skill: Diagnosing x-ray pictures.', 1988.

# Appendices

## Appendix A

## Consent Form

### PARTICIPANT CONSENT FORM

TAILORING A PHYSIOLOGICALLY DRIVEN RATING SYSTEM
MASTER THESIS BY HARRY VASANTH

**Purpose of this Study**
The purpose of the study is to evaluate and understand relevancy of physiological sensors in the context of gaming performance rating metrics.

**Procedures**
First of all, the participant will be wearing a composite sensor (smart watch) which measures heart rate using Photoplethysmography (PPG) sensors as well as spatial movement related measurements using the built-in accelerometer and gyroscope. Subsequently, the participant will be going through an introduction and initial warm up of an exergame. Afterwards, the participant will be facing several opponents (human participants) in a round robin tournament. After facing each opponent, the participant will be given a brief questionnaire to be answered.

**Participant Requirements**
You are eligible for participation if you are between 15-80 years old; present no conditions that may interfere with understanding, communication and the execution of the task; Understand English; Are motivated to participate.

Risks
The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life casual conversation.

**Benefits**
There may be no personal benefit from your participation in the study but the knowledge received may be of value to humanity.

Compensation & Costs
There will be a small compensation (food & drink) for your participation in this study. You may have transportation costs if you come from outside of UMa or M-ITI. There will be no additional cost to you if you participate in this study.

**Confidentiality**
By participating in the study, you understand and agree that Madeira Interactive Technologies Institute may be required to disclose your consent form, data and other personally identifiable information as required by law, regulation, subpoena or court order. Otherwise, your confidentiality will be maintained in the following manner:

Your data and consent form will be kept separate. Your consent form will be stored and will not be disclosed to third parties. By participating, you understand and agree that the data and information gathered during this study may be used by Madeira Interactive Technologies Institute and published and/or disclosed by Madeira Interactive Technologies Institute to others outside Madeira Interactive Technologies Institute. However, your name, e-mail and other direct personal identifiers in your consent form will not be mentioned in any such publication or dissemination of the research data and/or results. To protect your privacy, you will be assigned a code number and the collected data will be recorded by this code, NOT your name. Only the authorized researcher (Harry Vasanth) will have access to these data.

**Optional Permission**

# PARTICIPANT CONSENT FORM

TAILORING A PHYSIOLOGICALLY DRIVEN RATING SYSTEM
MASTER THESIS BY HARRY VASANTH

I understand that the researchers may want to use a short portion of any video recording and photographs for illustrative reasons in presentations of this work for scientific or educational purposes. I give my permission to do so.

Please initial here: _____YES_____NO

**Consent Form for Participation in Research**
Your participation is voluntary. You are free to stop your participation at any point. Refusal to participate or withdrawal of your consent or discontinued participation in the study will not result in any penalty or loss of benefits or rights to which you might otherwise be entitled. The researcher may at their discretion remove you from the study for any of a number of reasons. In such an event, you will not suffer any penalty or loss of benefits or rights which you might otherwise be entitled.

**Right to Ask Questions & Contact Information**
If you have any questions about this study, you should feel free to ask them now. If you have questions later please contact the lead researcher Harry Vasanth via harry.vasanth@m-iti.org

Voluntary Consent
By signing below, you agree that the above information has been explained to you and all your current questions have been answered. You are encouraged ask questions about any aspect of this study during the beginning and the end of the sessions as well as in the future. By signing this form, you agree to participate in this research.

_____ _____
PARTICIPANT SIGNATURE DATE

I certify that I have explained the nature and purpose of this research study to the above individual and I have discussed the potential benefits and possible risks of participation in the study. Any questions the individual has about this study have been answered and any future questions will be answered as they arise.

_____ _____
SIGNATURE OF PERSON OBTAINING CONSENT DATE

# Appendix B

## Flow Activity Experience Scale (DFS-2)

### Activity Experience Scale (DFS-2)

Please answer the following questions in relation to your experience in your chosen activity. These questions relate to the thoughts and feelings you may experience during participation in your activity. You may experience these characteristics some of the time, all of the time, or none of the time. There are no right or wrong answers. Think about how often you experience each characteristic during your activity and circle the number that best matches your experience.

**Rating scale**

| Never | Rarely | Sometimes | Frequently | Always |
|---|---|---|---|---|
| **1** | **2** | **3** | **4** | **5** |

*PLEASE CIRCLE ANSWER*

**When participating in _____ (name activity):**

1. I am challenged, but I believe my skills will allow me to meet the challenge.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

2. I make the correct movements without thinking about trying to do so.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

3. I know clearly what I want to do.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

4. It is really clear to me how my performance is going.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

5. My attention is focused entirely on what I am doing.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

6. I have a sense of control over what I am doing.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

7. I am not concerned with what others may be thinking of me.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

8. Time seems to alter (either slows down or speeds up).

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

9. I really enjoy the experience.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

10. My abilities match the high challenge of the situation.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

11. Things just seem to happen automatically.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

12. I have a strong sense of what I want to do.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

13. I am aware of how well I am performing.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

14. It is no effort to keep my mind on what is happening.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

15. I feel like I can control what I am doing.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

16. I am not concerned with how others may be evaluating me.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

17. The way time passes seems to be different from normal.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

***CONTINUES OVER***

**Rating scale**

| Never | Rarely | Sometimes | Frequently | Always |
|---|---|---|---|---|
| **1** | **2** | **3** | **4** | **5** |

*PLEASE CIRCLE ANSWER*

**When participating in** _____ **(name activity):**

**18.** I love the feeling of the performance and want to capture it again.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**19.** I feel I am competent enough to meet the high demands of the situation.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**20.** I perform automatically, without thinking too much.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**21.** I know what I want to achieve.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**22.** I have a good idea while I am performing about how well I am doing.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**23.** I have total concentration.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**24.** I have a feeling of total control.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**25.** I am not concerned with how I am presenting myself.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**26.** It feels like time goes by quickly.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**27.** The experience leaves me feeling great.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**28.** The challenge and my skills are at an equally high level.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**29.** I do things spontaneously and automatically without having to think.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**30.** My goals are clearly defined.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**31.** I can tell by the way I am performing how well I am doing.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**32.** I am completely focused on the task at hand.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**33.** I feel in total control of my body.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**34.** I am not worried about what others may be thinking of me.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**35.** I lose my normal awareness of time.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

**36.** The experience is extremely rewarding.

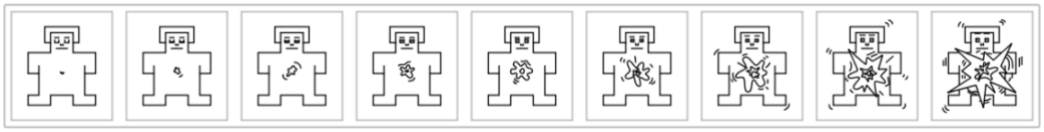
| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|

# Appendix C

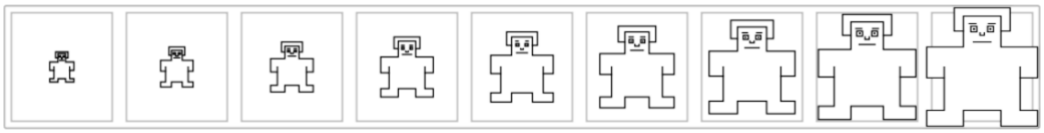
## The Self-Assessment Manikin

Please rate your experience using the following scales:

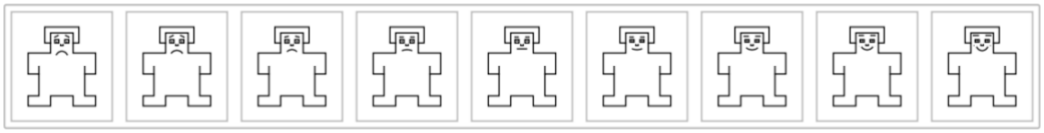
Negative Neutral Positive



Negative Neutral Positive



Negative Neutral Positive



# Appendix D

## Positive and Negative Affect Schedule (PANAS-SF)

Positive and Negative Affect Schedule (PANAS-SF)

| | | Very slightly or not at all | A little | Moderately | Quite a bit | Extremely |
|---|---|---|---|---|---|---|
| PANAS 1 | Interested | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 2 | Distressed | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 3 | Excited | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 4 | Upset | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 5 | Strong | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 6 | Guilty | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 7 | Scared | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 8 | Hostile | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 9 | Enthusiastic | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 10 | Proud | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 11 | Irritable | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 12 | Alert | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 13 | Ashamed | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 14 | Inspired | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 15 | Nervous | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 16 | Determined | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 17 | Attentive | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 18 | Jittery | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 19 | Active | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |
| PANAS 20 | Afraid | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 |

## Appendix E

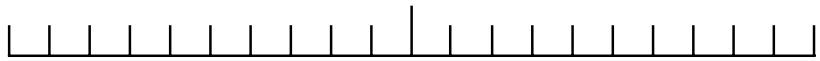
## NASA Task Load Index (TLX)

### *NASA Task Load Index*

*Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.*

| Name | Task | Date |
|---|---|---|
| | | |

Mental Demand — How mentally demanding was the task?



Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

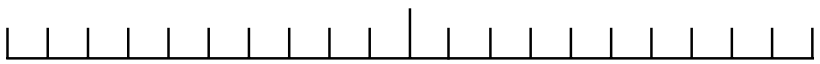
Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?



Very Low — Very High

## Appendix F

### BORG Rating of Perceived Exertion (BORG-RPE)

| | |
|---|---|
| 6 | No exertion at all |
| 7 | Extremely light |
| 8 | |
| 9 | Very light |
| 10 | |
| 11 | Light |
| 12 | |
| 13 | Somewhat hard |
| 14 | |
| 15 | Hard (heavy) |
| 16 | |
| 17 | Very hard |
| 18 | |
| 19 | Extremely hard |
| 20 | Maximal exertion |

## Appendix G

## Participants Information

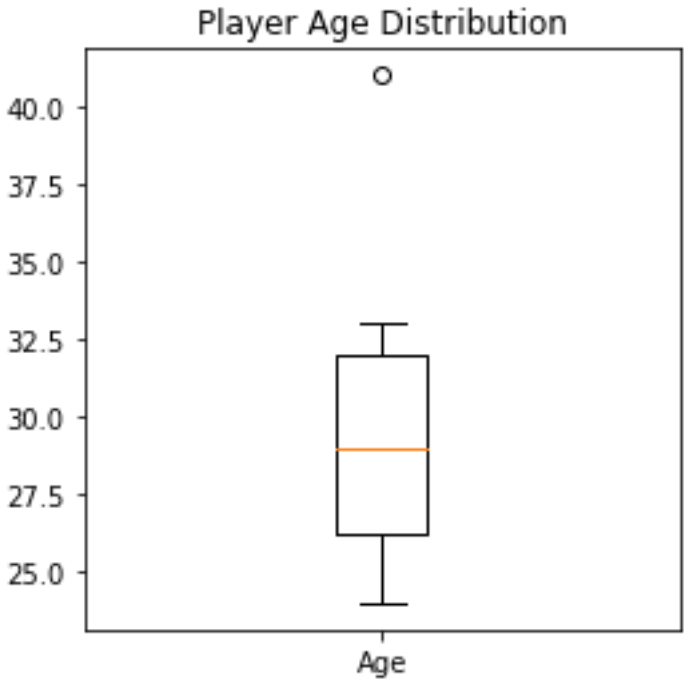



Figure 30. Player age distribution

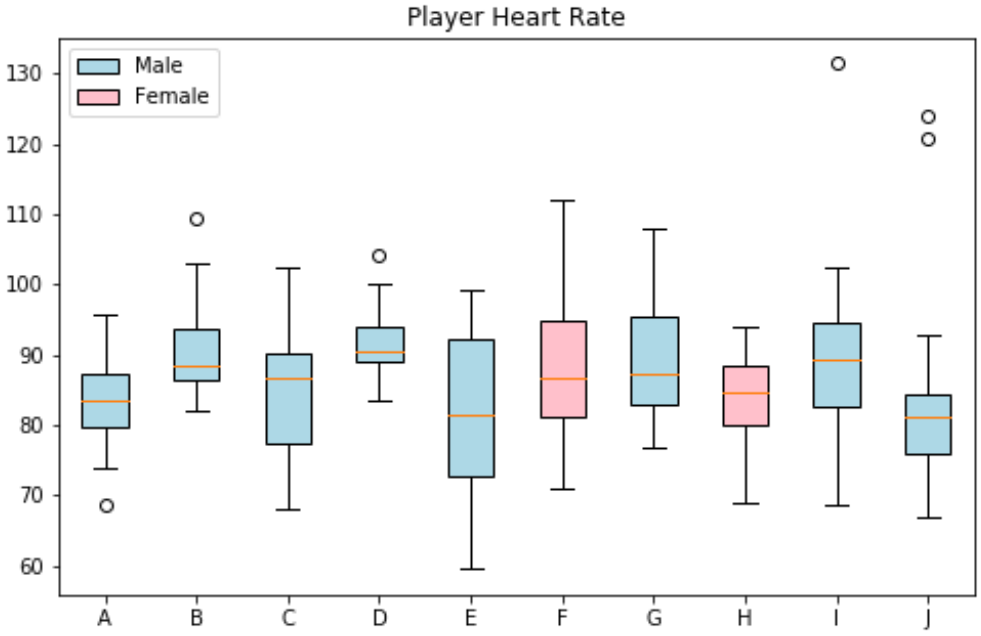



Figure 31. Player baseline heart rate and gender distribution

# Appendix H

## Individual Correlations for each Player

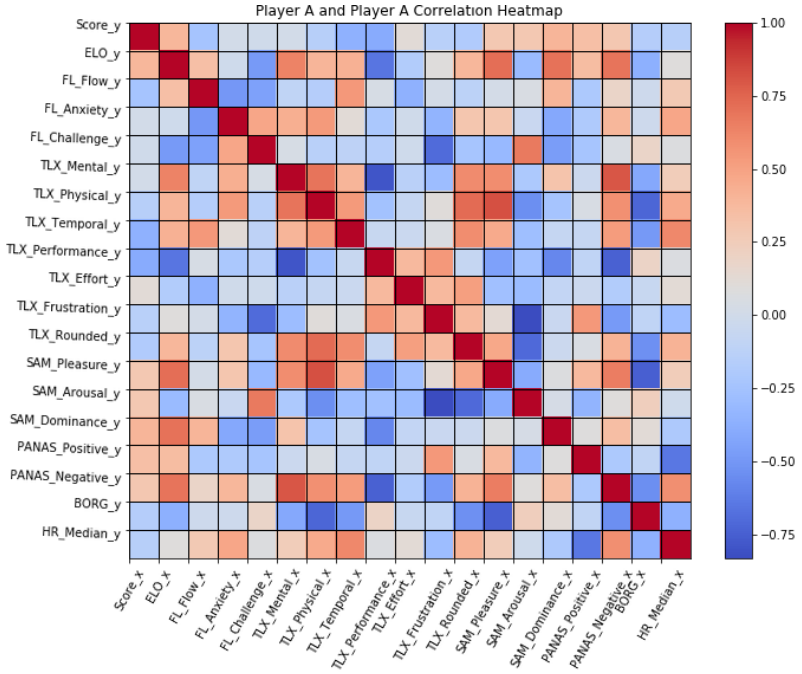



Figure 32. Player A Correlations

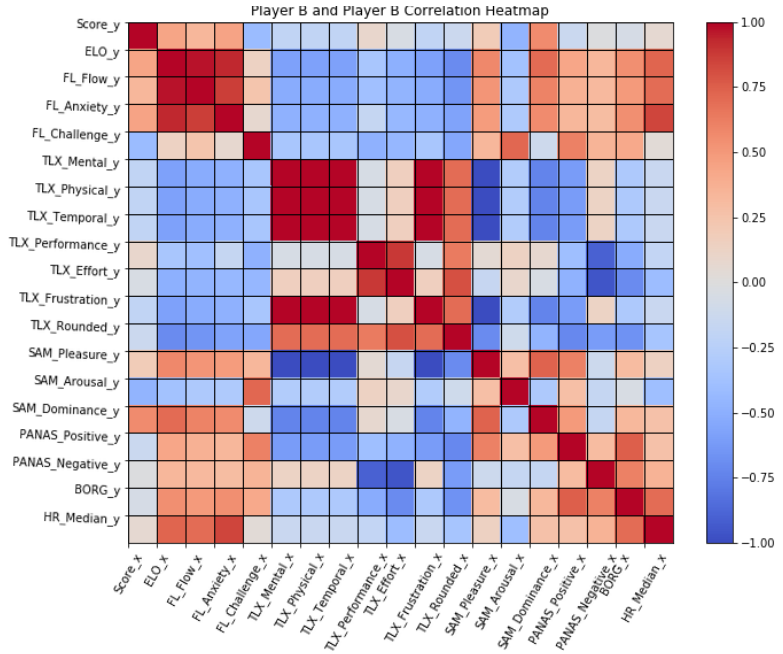



Figure 33. Player B Correlations

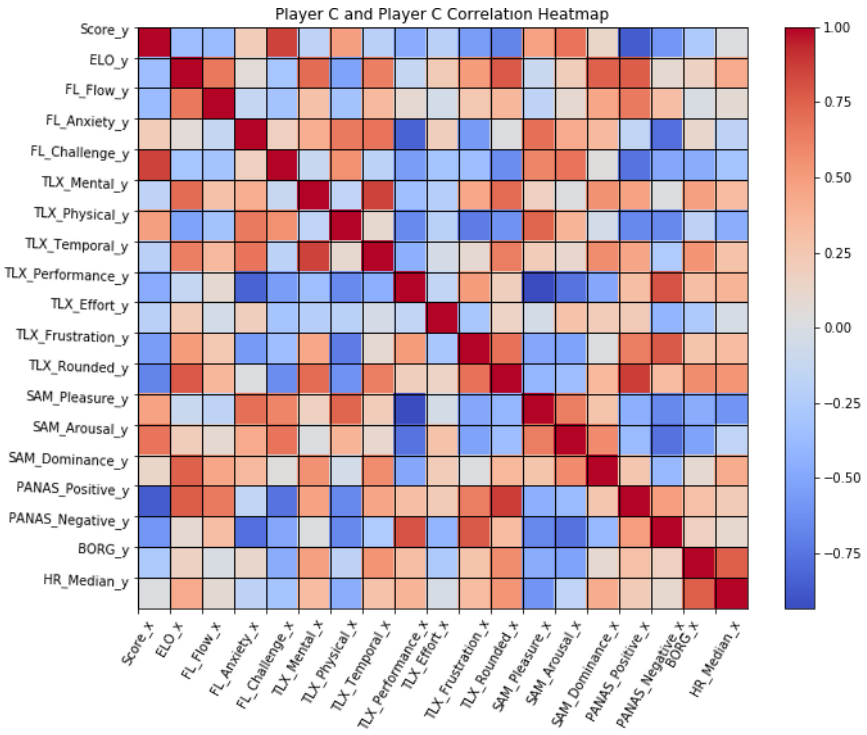



Figure 34. Player C Correlations

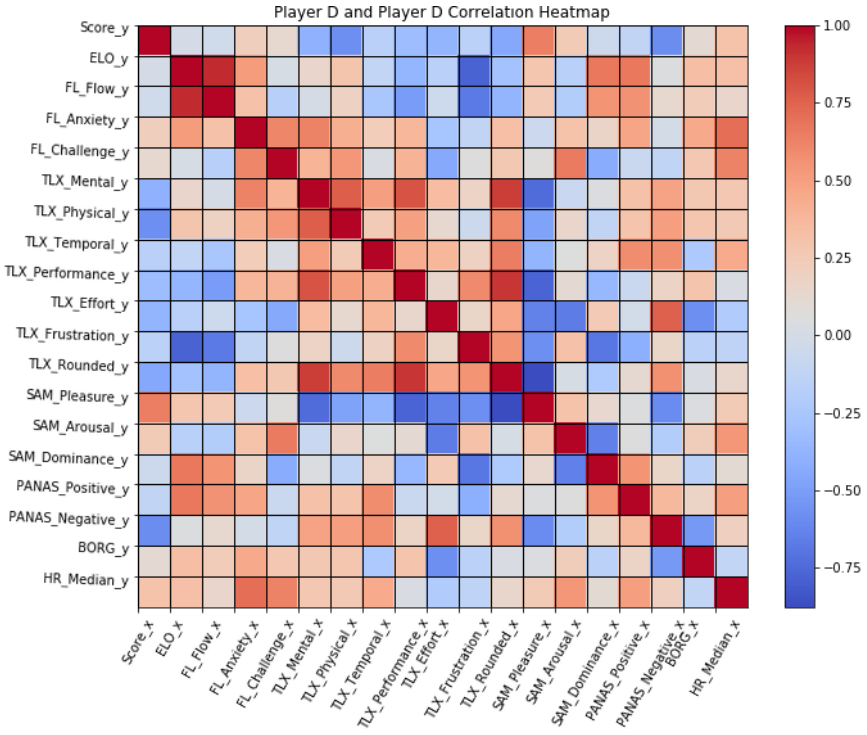



Figure 35. Player D Correlations

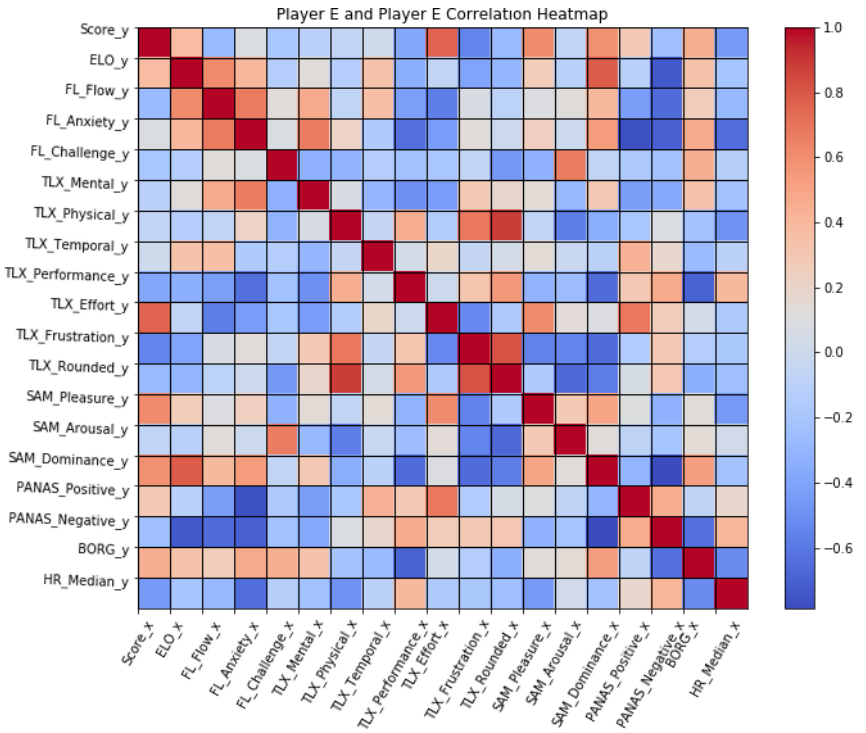



Figure 36. Player E Correlations

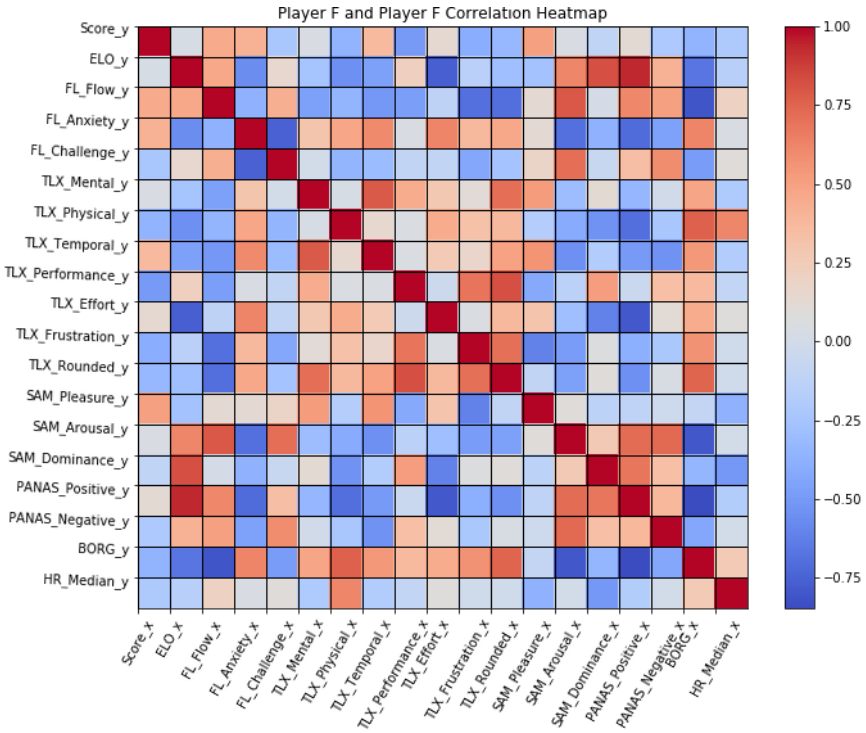



Figure 37. Player F Correlations

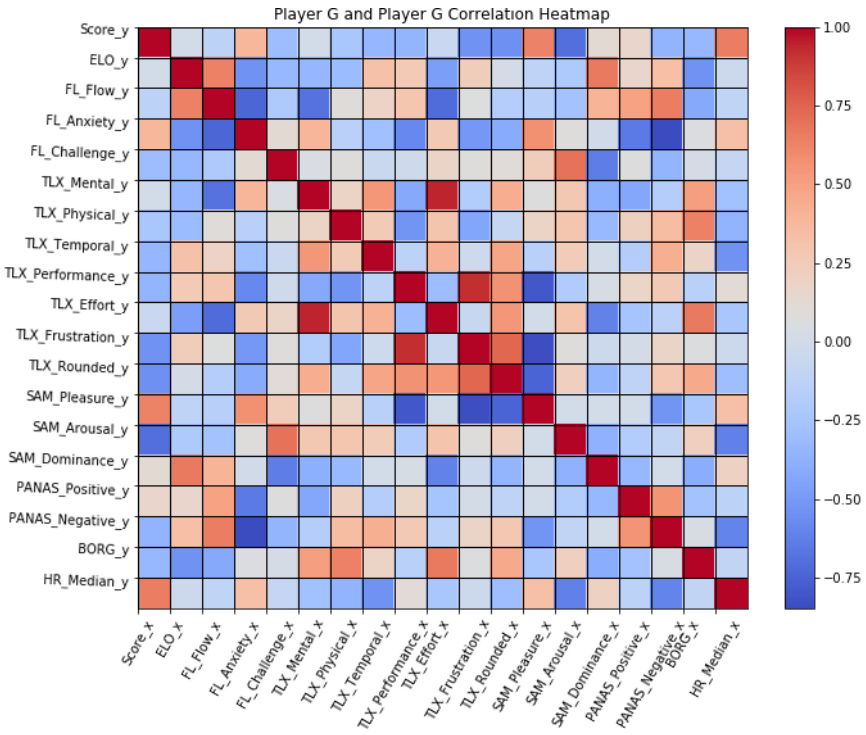



Figure 38. Player G Correlations

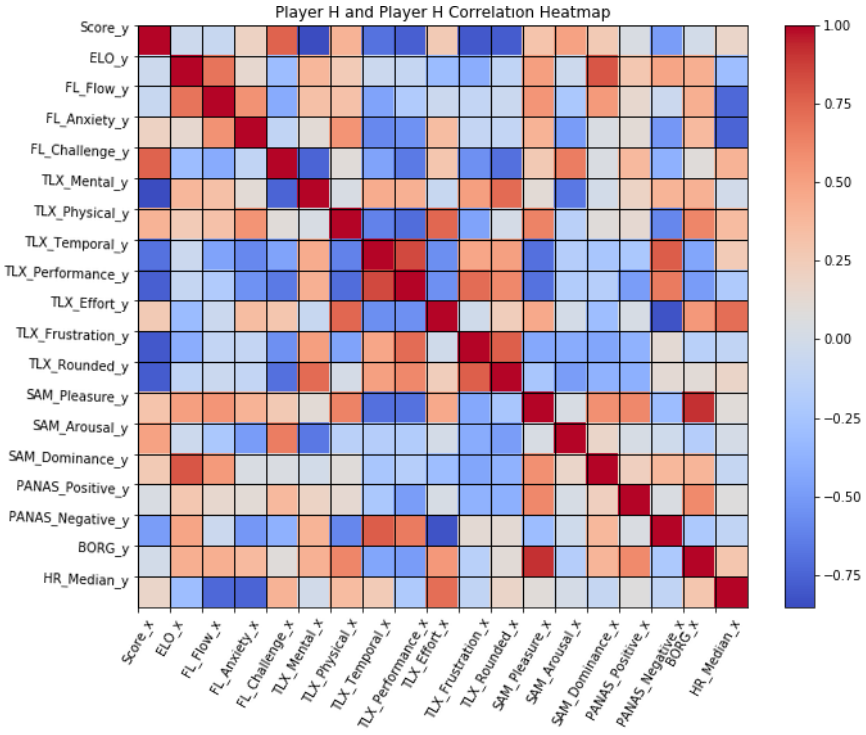



Figure 39. Player H Correlations

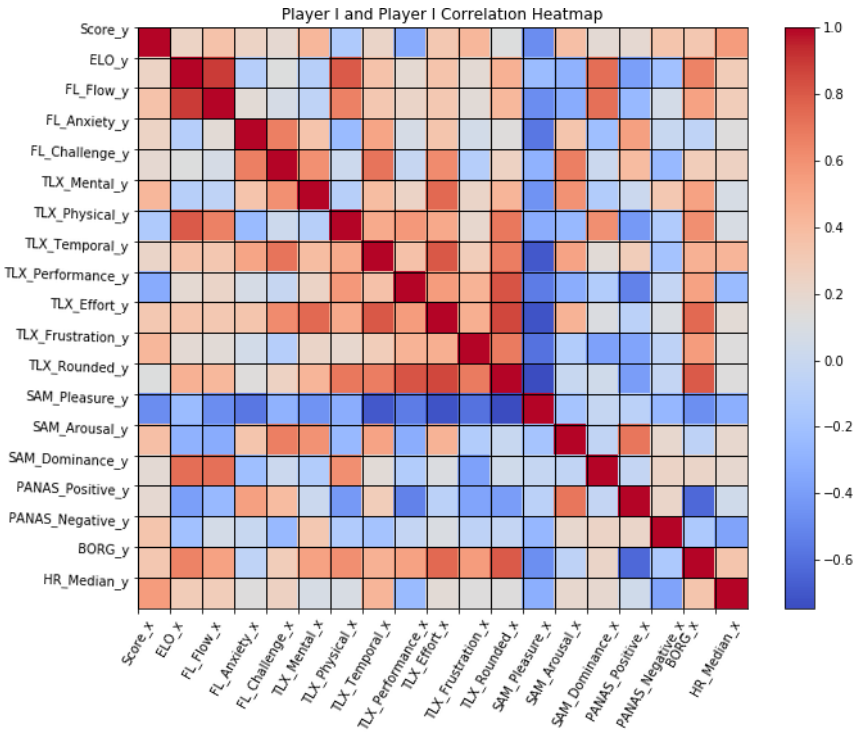



Figure 40. Player I Correlations

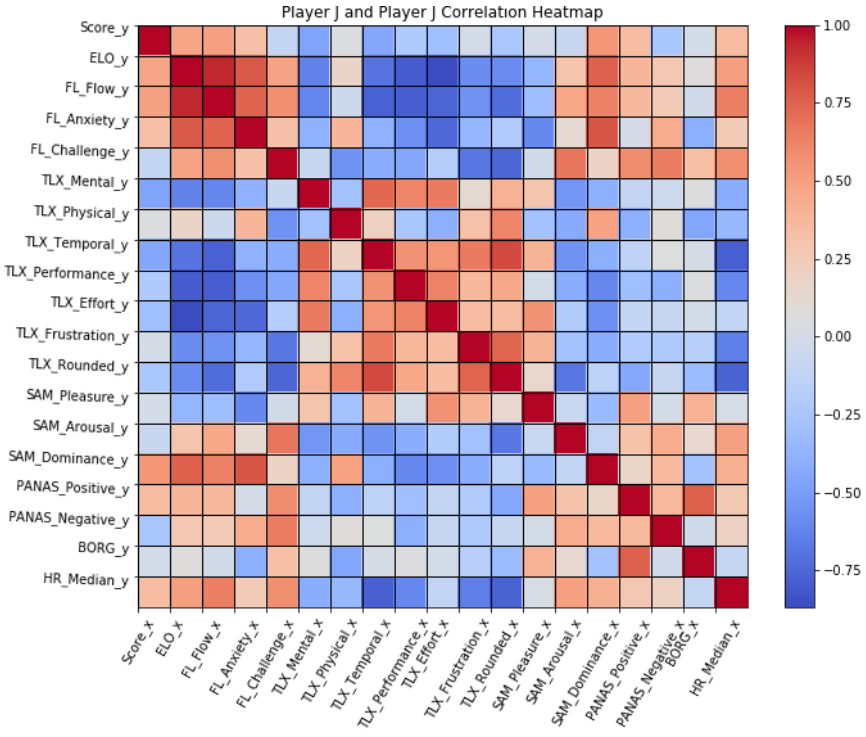



Figure 41. Player J Correlations

# Appendix I

## Progression of instrument variables throughout the tournament

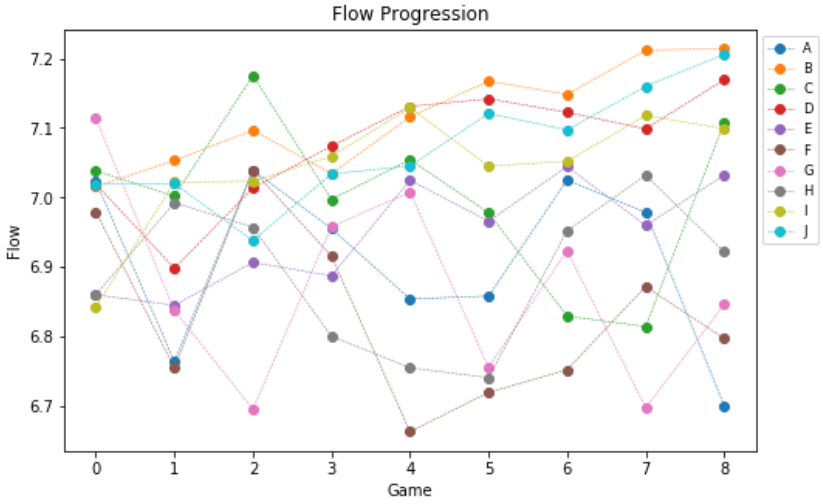



Figure 42. Flow progression

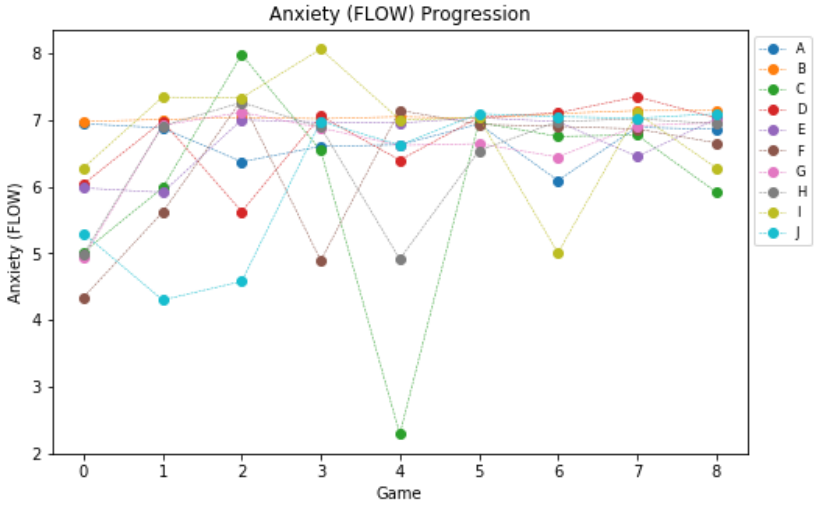



Figure 43. Anxiety (FLOW) Progression

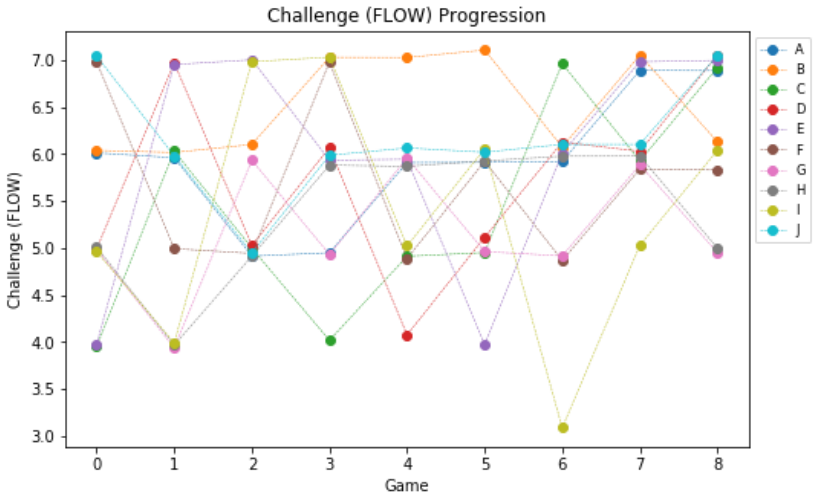



Figure 44. Challenge (FLOW) Progression

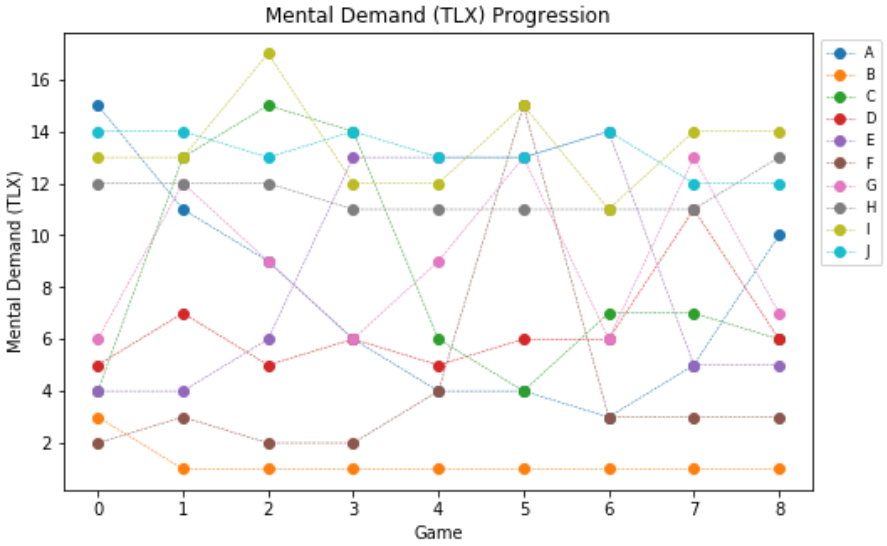



Figure 45. Mental Demand (TLX) Progression

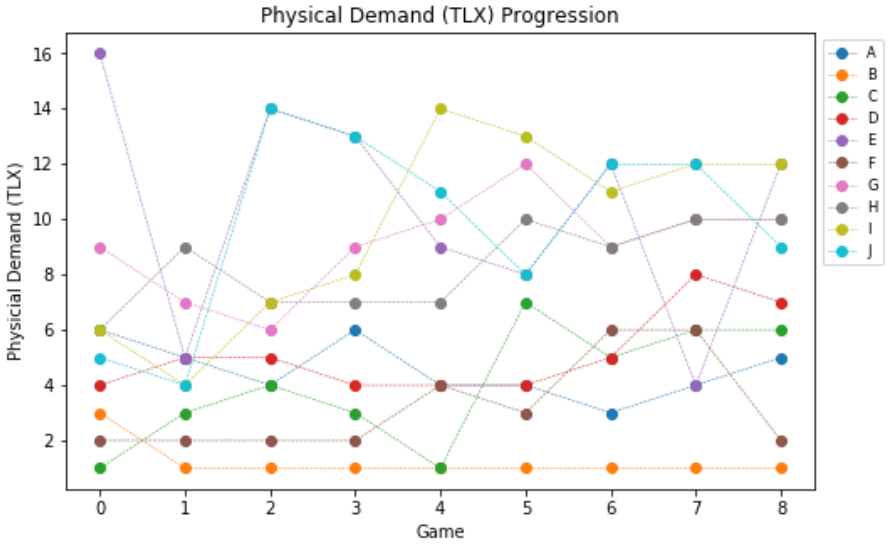



Figure 46. Physical Demand (TLX) Progression

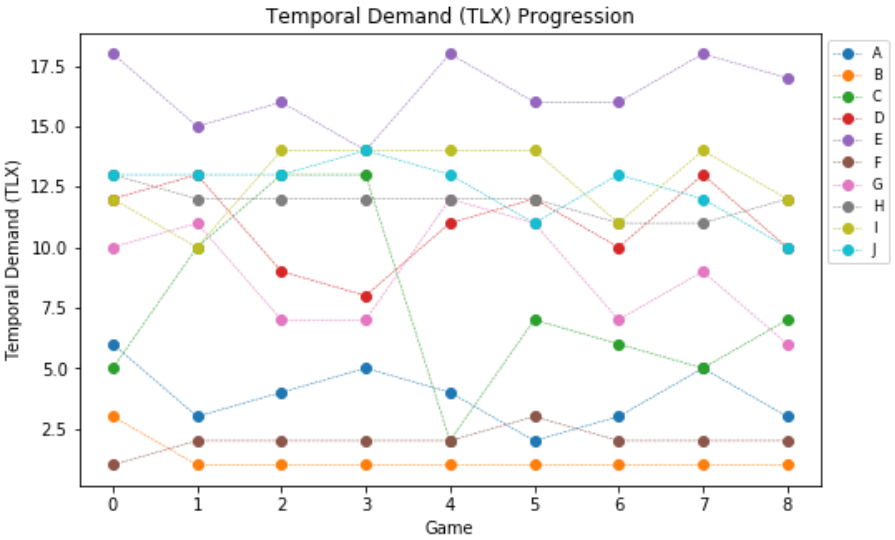



Figure 47. Temporal Demand (TLX) Progression

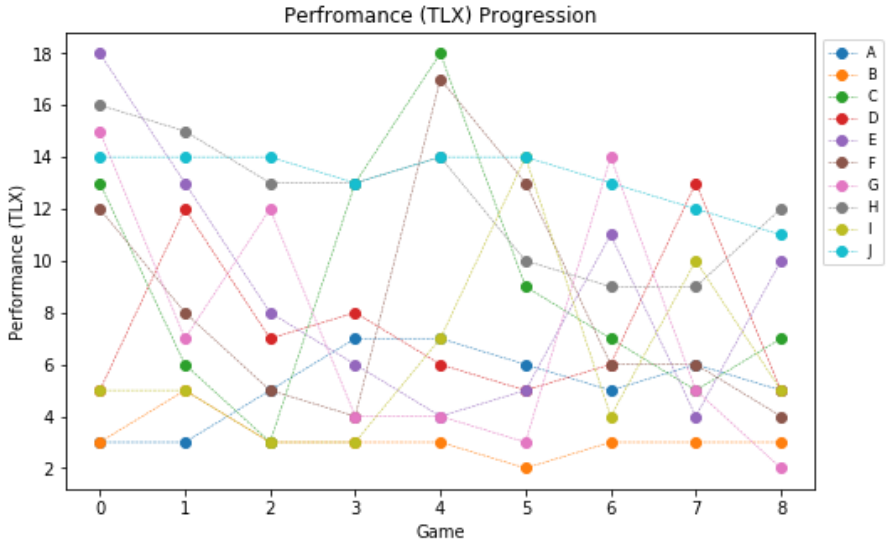



Figure 48. Performance (TLX) Progression

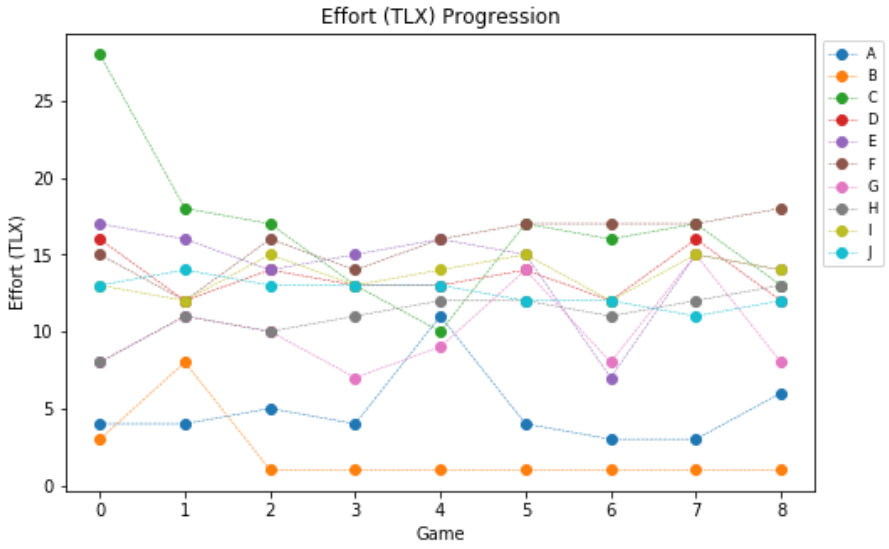



Figure 49. Effort (TLX) Progression

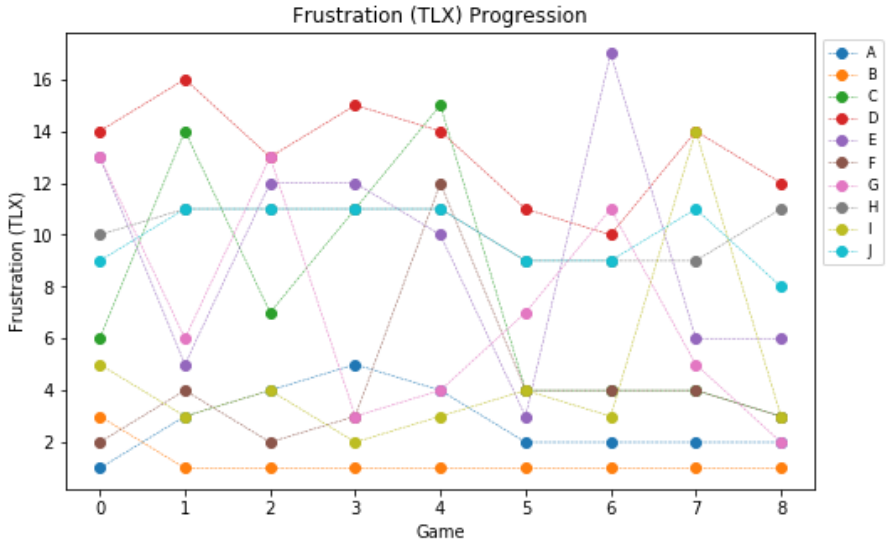



Figure 50. Frustration (TLX) Progression

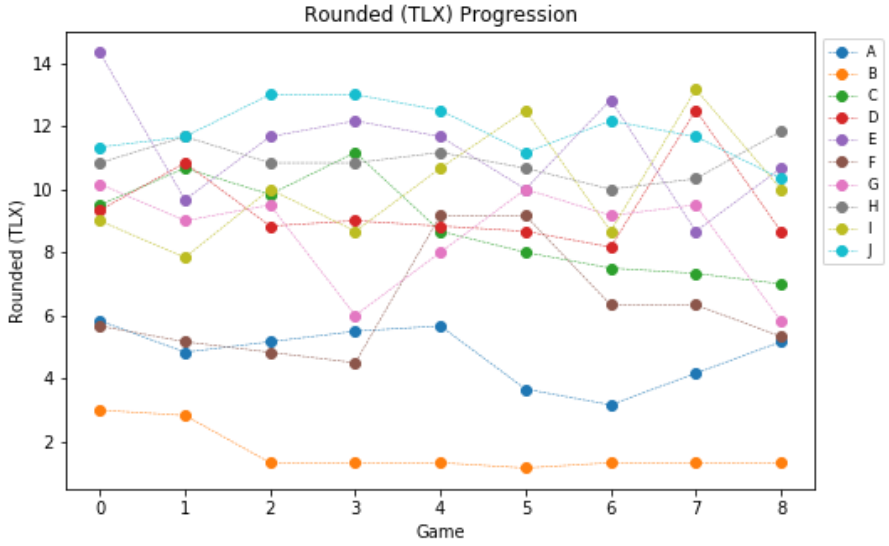



Figure 51. Rounded TLX Progression

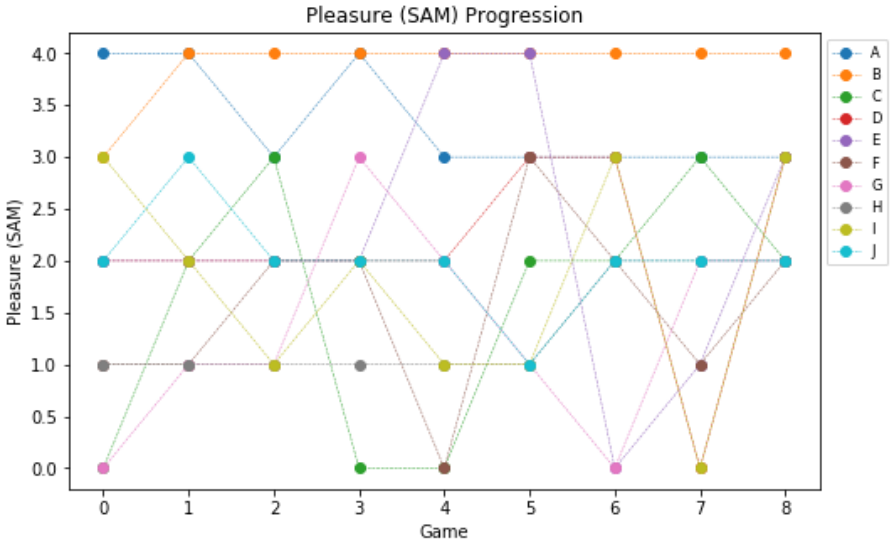



Figure 52. Pleasure (SAM) Progression

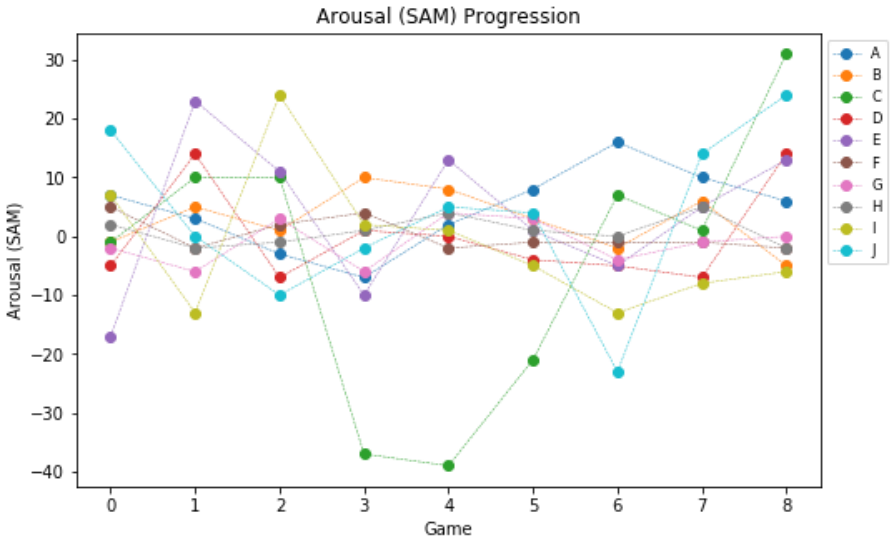



Figure 53. Arousal (SAM) Progression

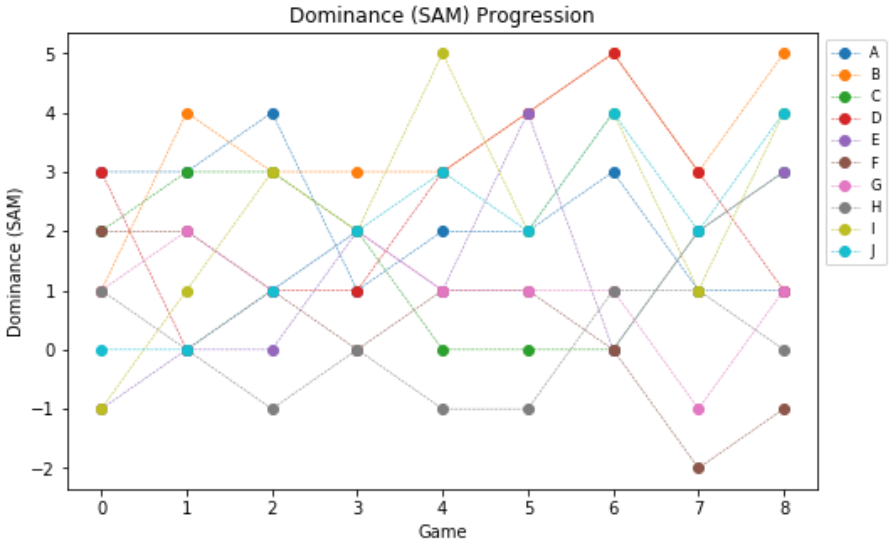



Figure 54. Dominance (SAM) Progression

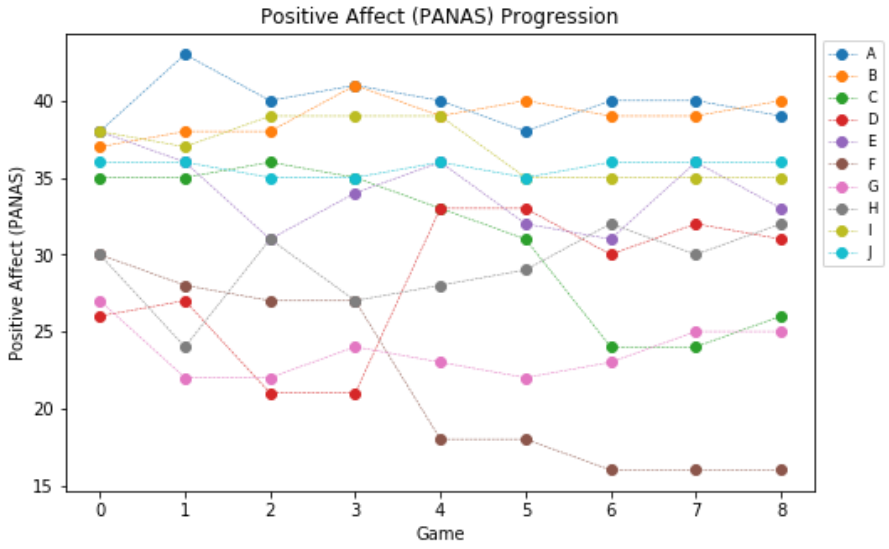



Figure 55. Positive Affect (PANAS) Progression

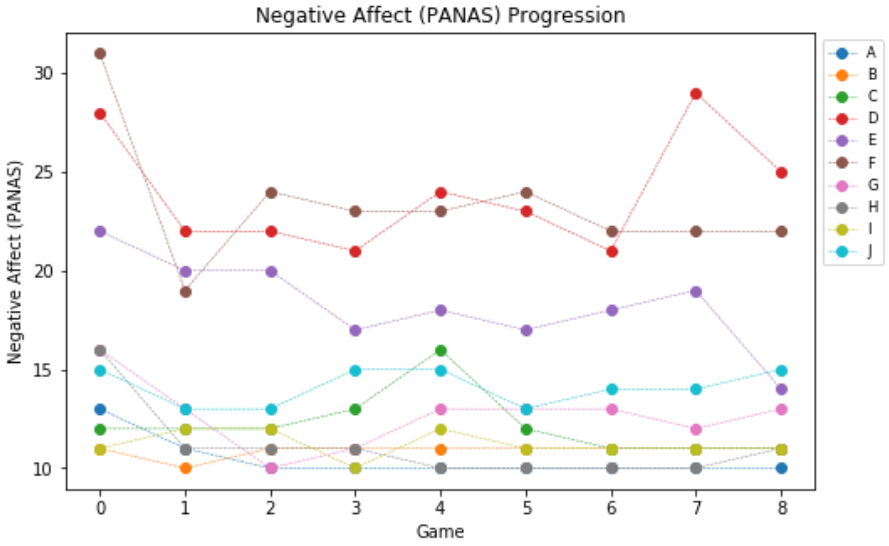



Figure 56. Negative Affect (PANAS) Progression

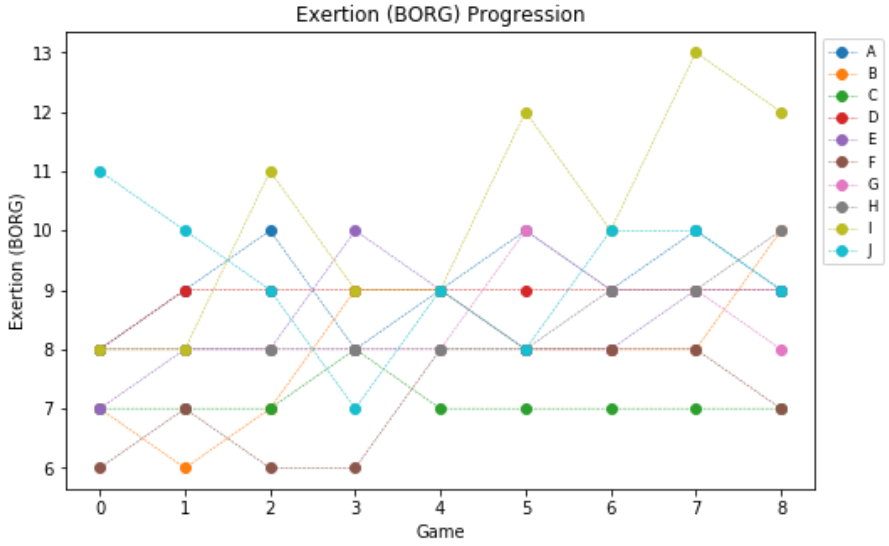



Figure 57. Exertion (BORG) Progression

# Appendix J

## Regressions of instrument variables for ELO

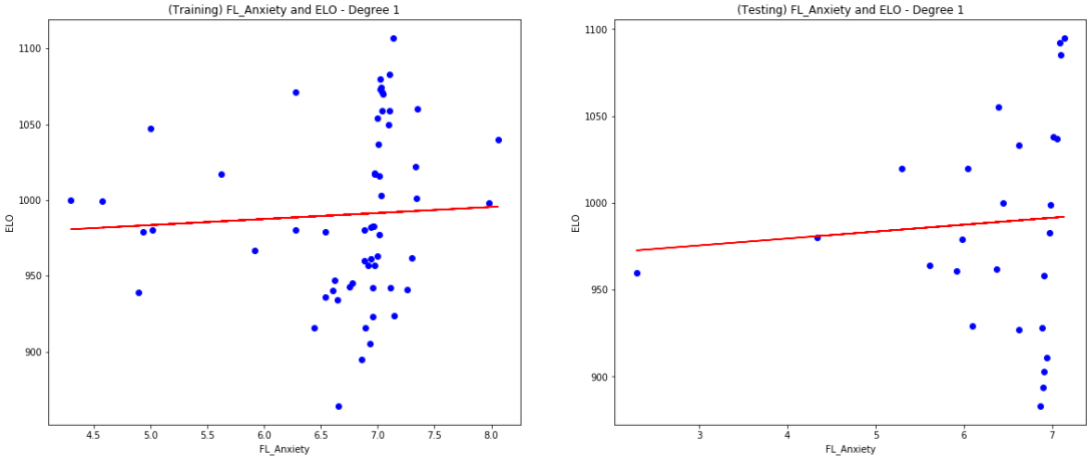



Figure 58. Anxiety (FLOW) Training and Testing

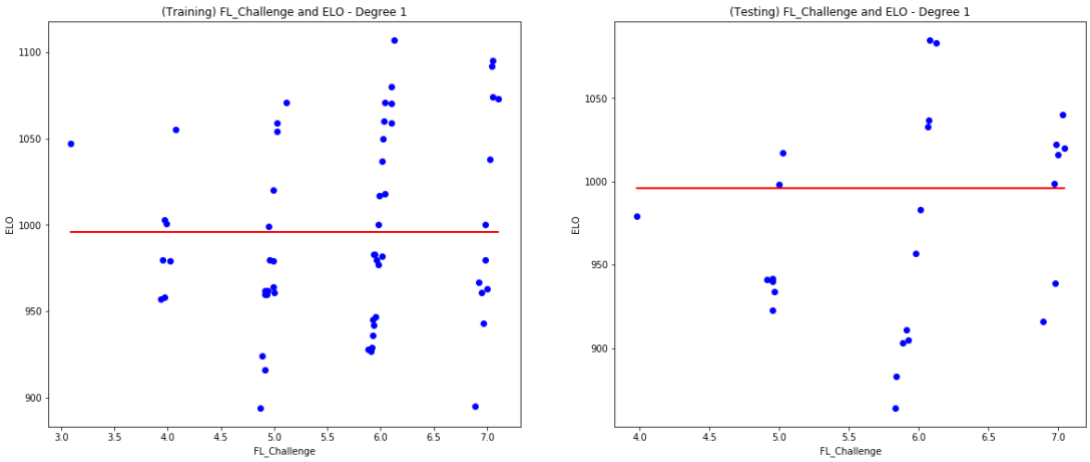



Figure 59. Challenge (FLOW) Training and Testing

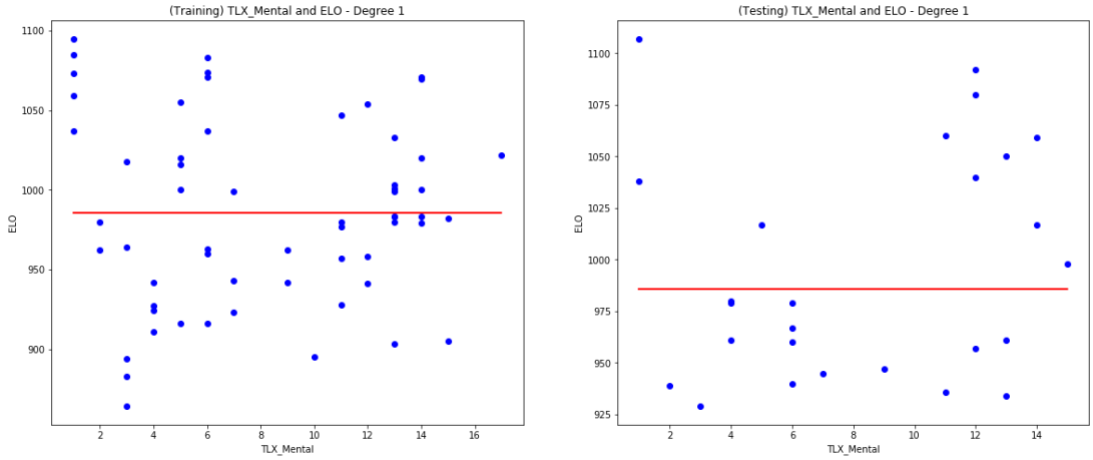



Figure 60. Mental Demand (TLX) Training and Testing

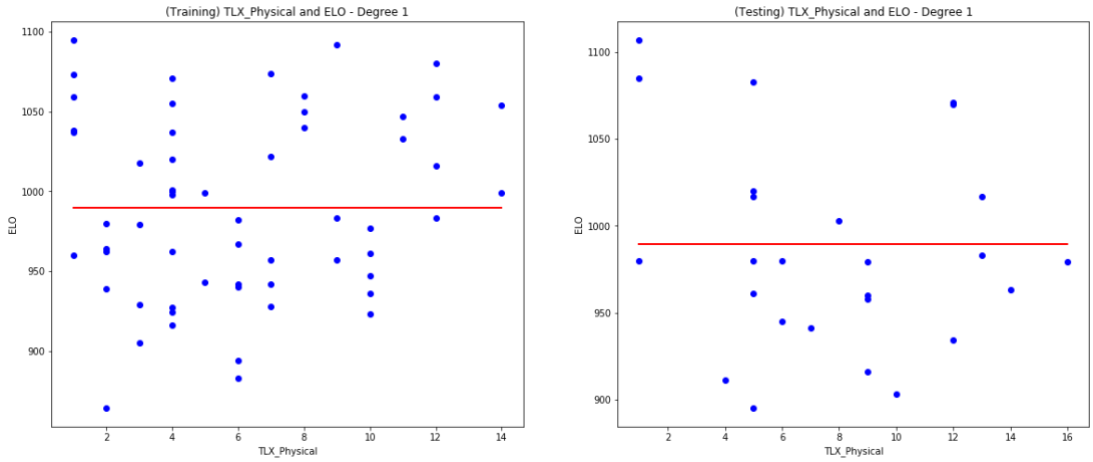



Figure 61. Physical Demand (TLX) Training and Testing

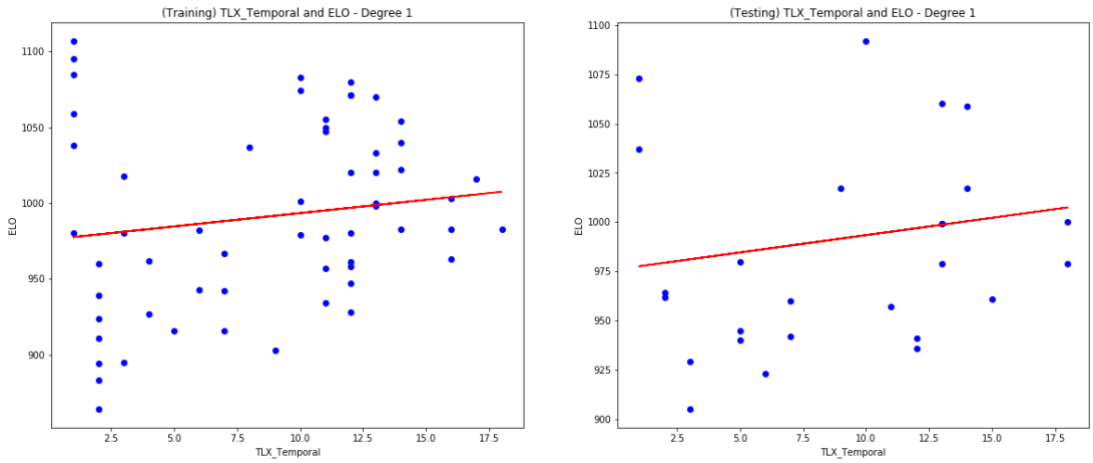



Figure 62. Temporal Demand (TLX) Training and Testing

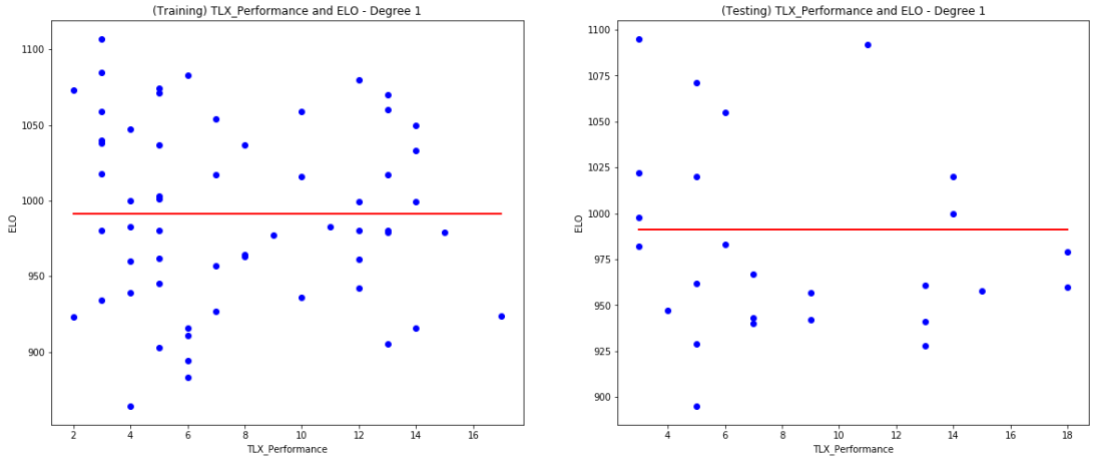



Figure 63. Performance (TLX) Training and Testing

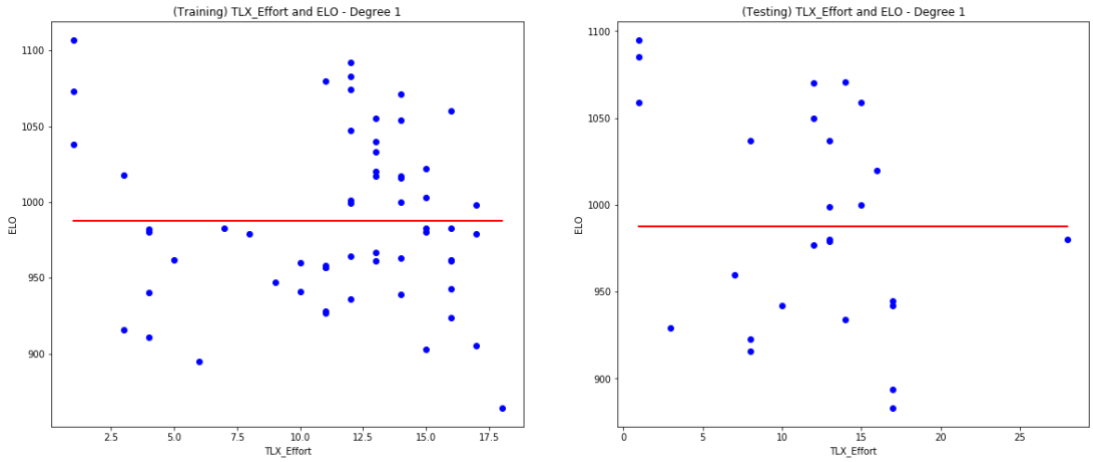



Figure 64. Effort (TLX) Training and Testing

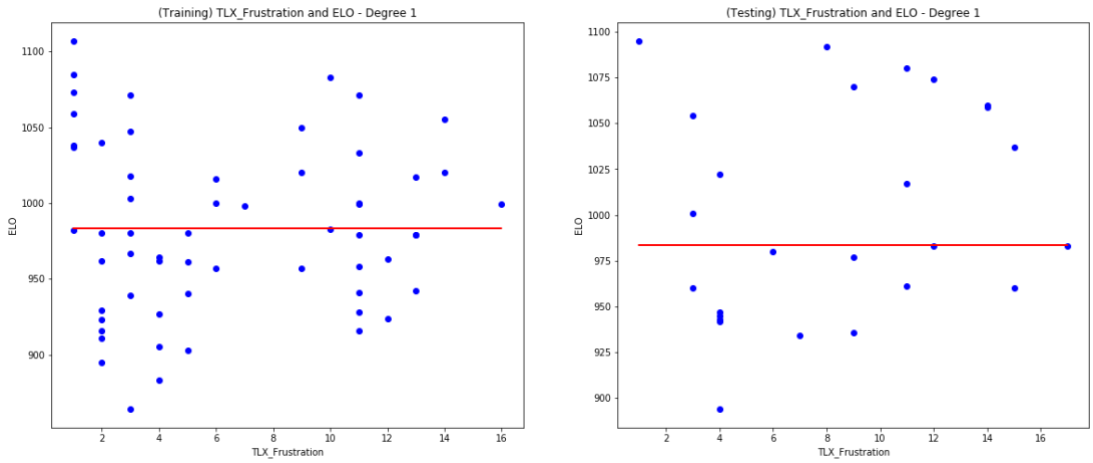



Figure 65. Frustration (TLX) Training and Testing

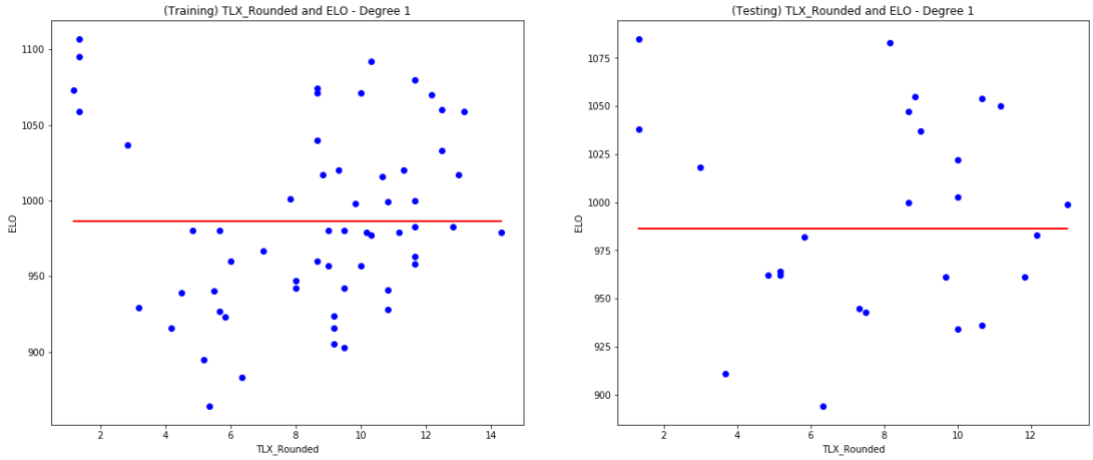



Figure 66. Rounded TLX Training and Testing

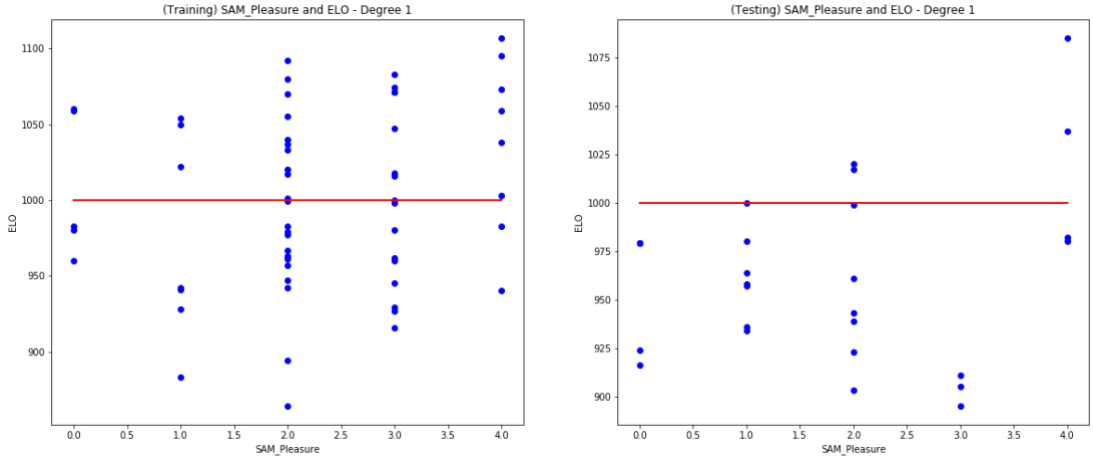



Figure 67. Pleasure (SAM) Training and Testing

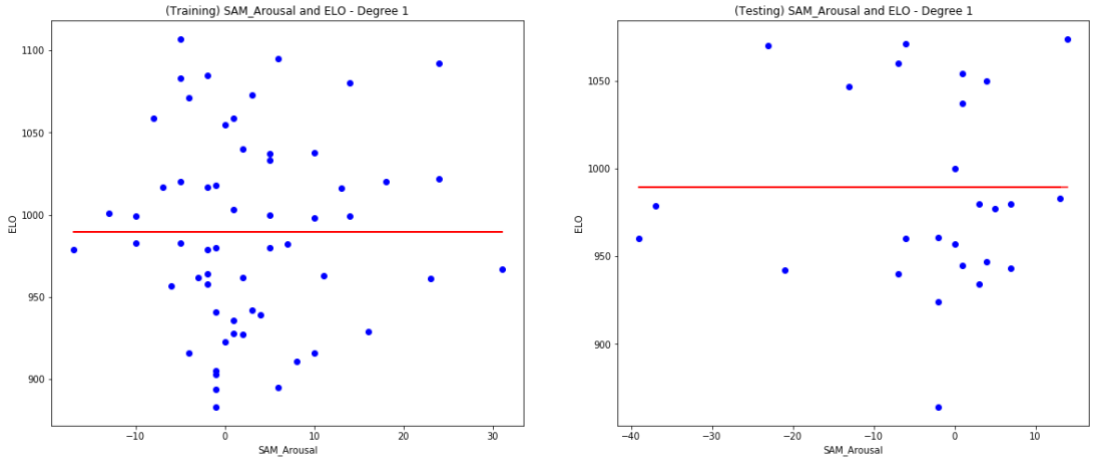



Figure 68. Arousal (SAM) Training and Testing

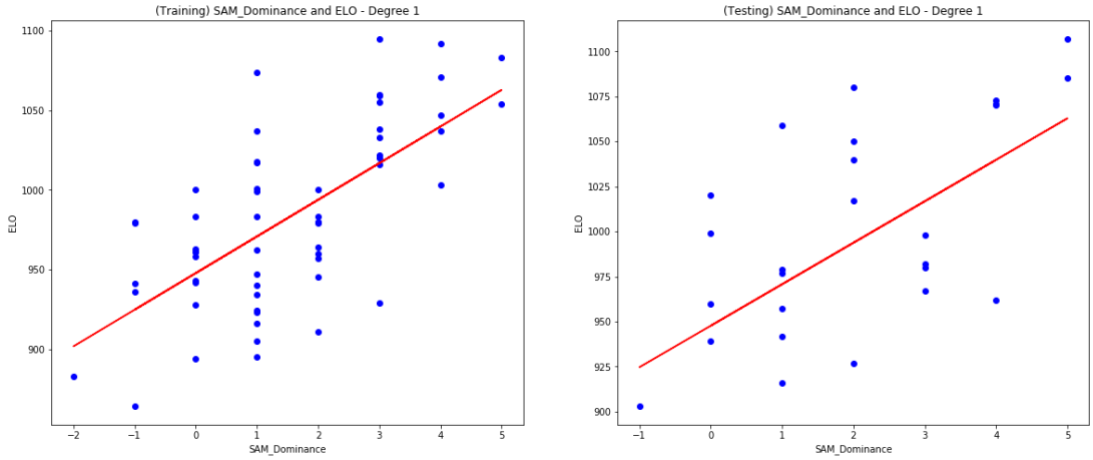



Figure 69. Dominance (SAM) Training and Testing

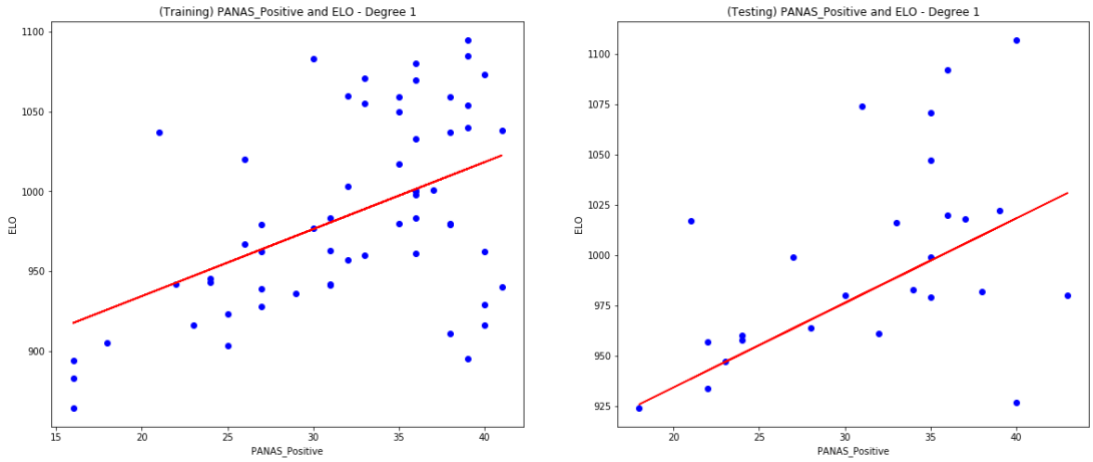



Figure 70. Positive Affect (PANAS) Training and Testing

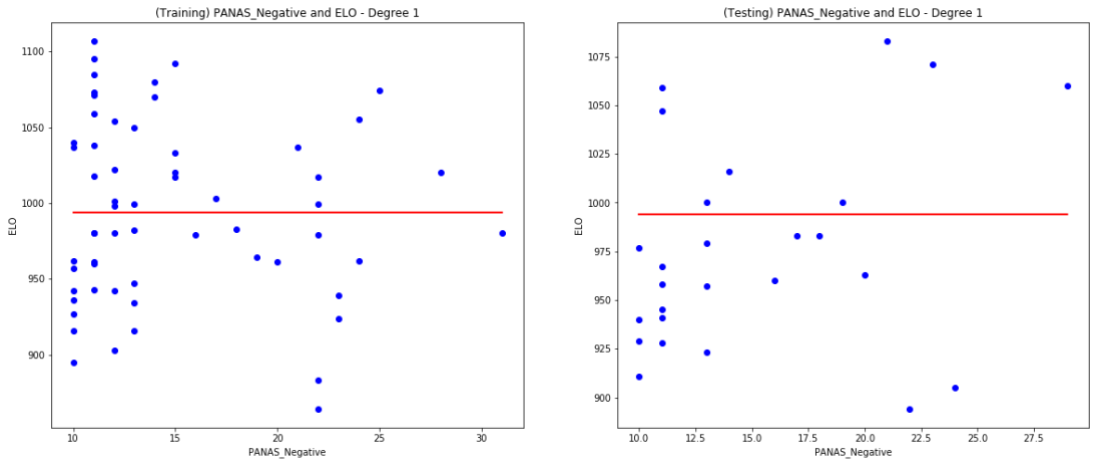



Figure 71. Negative Affect (PANAS) Training and Testing

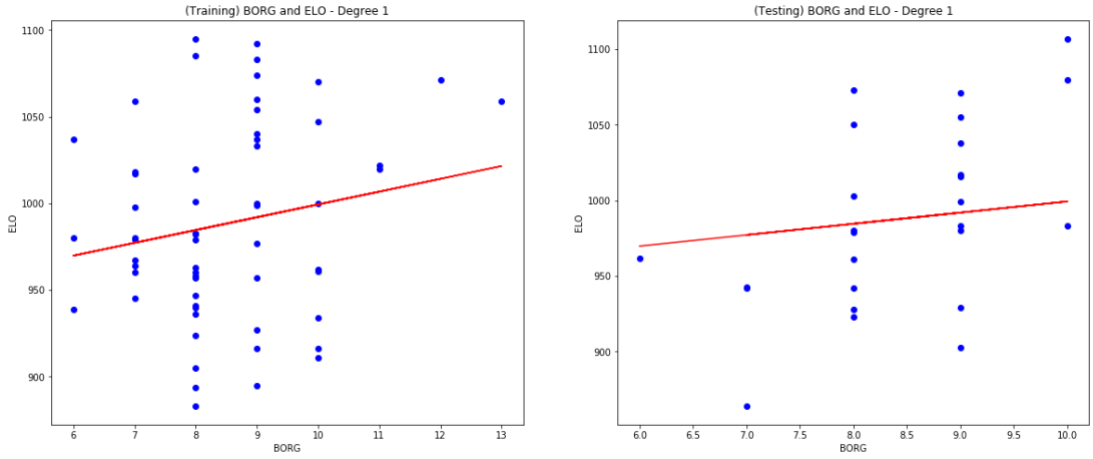



Figure 72. Exertion (BORG) Training and Testing

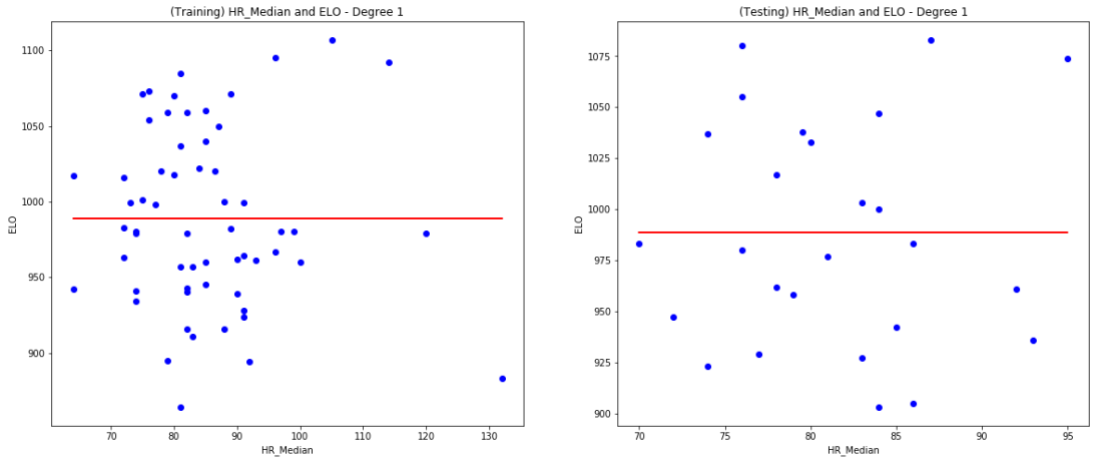



Figure 73. HR Median Training and Testing