The effects of gender-segregated competition on learning and performance in chess

Jorge Martínez Ramírez

July 22, 2021

Author:

Jorge Alberto Martínez Ramírez
jorgea.martinezr@utadeo.edu.co

Director:

Andrés Fernando Bernal
andresf.bernale@utadeo.edu.co

Degree

Masters degree in Modeling and Simulation (Research)

Contents

Introduction

The difference between men and women's ratings in chess is notable. In an activity in which the demand is not on physical but on mental performance, this difference is still in need of a broadly accepted explanation (Bilalić, Smallbone, et al. 2009; Knapp 2010). Not only are women underrepresented at high ratings compared to the female-to-male ratio in the general population (Figure 1 left), but also, as higher ratings are examined, women become even less represented (Figure 1 right).

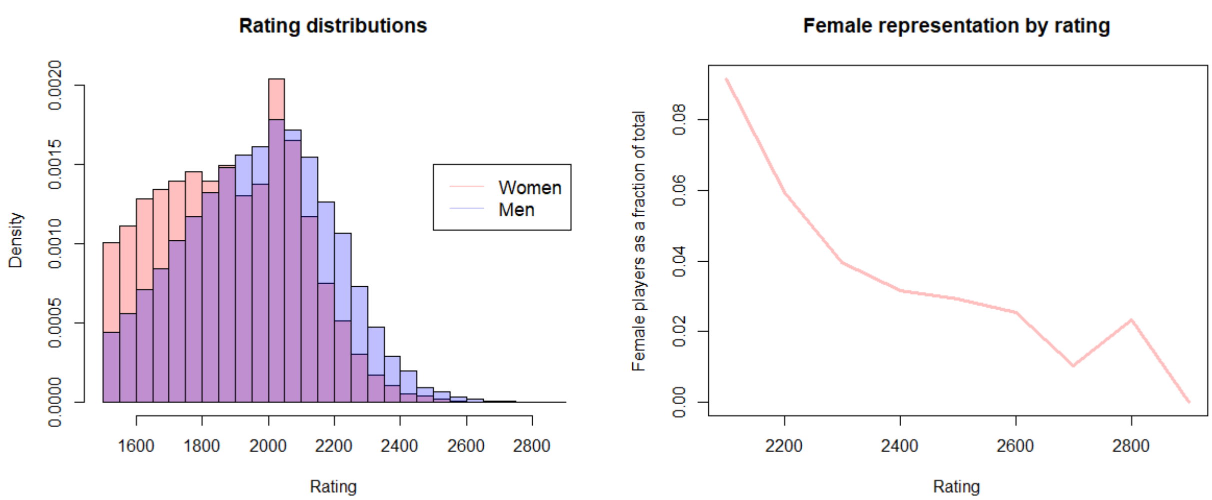

Figure 1: Men and women's rating distributions (left) and female representation by rating (right) in the December 2012 FIDE rating list.

One salient feature of the differences between men and women, specially at the elite level, is its persistence in time. Despite the efforts to attract more women to the game and the growth in the number of women with Grand Master titles in later years, rating differences between the best men and the best women have persisted (Howard 2005).

One of the efforts that have been taken to support the development of women's chess is the creation of women-only tournaments. From the perspective of part of the chess world, the existence of these exclusively female tournaments is vital for the continuation of female participation in chess, as well as to attract new female players to the circuit (Phillips 2010; Polgar 2019; Shahade 2005). However, some of the best female players in history have opted out of women-only tournaments and top female chess players have spoken against these events arguing that they are condescending to women and an obstacle to the improvement of female chess (Phillips 2010; Shahade 2005).

This last point suggests some questions: Could women-only tournaments play a role in the sex gap seen in the elite of chess? Is it possible, for instance, that the top women aren't playing as much as they could against opponents that are better than them in games and tournaments, from which they could learn more and gain more experience?

With these ideas in mind, this work explores the effect that players' participation in female-only tournaments has on learning and performance. To achieve this, we formulated a mathematical model of the effect of learning from the challenges faced in competitive play. This model incorporates the concept of Desirable Difficulties (E. L. Bjork, R. Bjork, and Mcdaniel 2011; R. A. Bjork 1994). It was implemented in an agent-based model that we analysed on different scenarios with parameters that make sense according to actual chess competition data.

The central question studied in this work was: *What is the effect of players' participation in female-only tournaments on the distribution of ratings in chess?* In particular, we were interested in the performance of the elite female players and establishing if the preference that these players have for these segregated tournaments has any effect on how good the best players of that group can be. In a more general sense, the objective of this work was to establish the relationship between group segregation and performance in a competitive environment from the perspective of the learning opportunity that may be lost when a group - female players in the case of chess - restricts to playing mostly among themselves. It was raised from the intuition that a learning opportunity is lost when a player refrains from the challenge of facing harder opponents, as would be the case when women refrain from playing against the higher-ranked male population. Research from the learning sciences supporting this intuition will be explained in the conceptual framework section of this work.

In this sense, our hypothesis is that self-segregating within a group that has a lower mean measure of performance - lower rating - and lower participation rate at higher ratings has a negative effect on the rating improvement of that group's best players. In the world of chess, this would imply that a generalized preference by women to participate in women-only tournaments would be in detriment of women's rating improvement in the elite levels.

Our goal is to contribute to the understanding of observed differences between the sexes in an intellectual competitive activity in which many of the explanations offered by previous studies are possibly only a piece of a bigger puzzle. The fact is that men and women don't compete fully integrated in chess and, as we will explain, there is reason to suspect that this is a factor that can't be dismissed as part of an explanation without being studied.

On the other hand, this work is by no means an attempt to propose a single, ultimate explanation to a complex problem, but an attempt to give insight into a phenomenon that must be taken into account: the possible learning cost of self-segregation. It is also not an attempt to asses the value that women-only tournaments may have from perspectives different to the learning cost examined here. It may very well be the case that women-only tournaments are worth keeping for valid reasons that are completely unrelated to the ones discussed in this work, even if some levels of segregation did make improvement more difficult.

The understanding of a disparity such as the one found between sexes in chess may help with the development and promotion of women's participation in the game. From a social perspective, it is also important to note that the misunderstanding of the causes for the rating differences in chess makes it more possible for these differences to be used as arguments that reinforce stereotypes, harming the broader conversation about gender equity.

Finally, the intention of this work is to use mathematical modeling theories and techniques to contribute to the scientific understanding of a complex social system.

With this goals in mind, in this document we begin by presenting in Chapter 1 an overview of the relevant concepts and theoretical frameworks used to address the research question and objectives. In Chapter 2 we perform a statistical study of a sample of FIDE-rated games, in order to assess the reality of self-segregation in female chess and its correlation with the rating gap. Then, in Chapter 3 we present a formulation of an Agent-based model to test the plausibility that segregation plays a role in shaping group differences in performance through a learning mechanism. In Chapter 4 we present and discuss the results from simulations of the Agent-based model. Then, in Chapter 5, we analyse the results from multiple simulations with varying model parameters to explore how the specific choice of parameters can affect the observed results. Finally, we present our general Conclusions.

Chapter 1

Conceptual Framework

1.1 Desirable difficulties in learning

Learning can be understood as the acquisition by an individual of new understandings and abilities that can later be recalled and sustained in the long-term. During training, or any learning-oriented activity, the performance of the learner is the collection of behaviors that can be measured and observed related to what is currently being learned. Learning, in contrast with performance, has to do with long-term changes and cannot be readily observed, it can only be inferred during training or instruction (Soderstrom and Robert A. Bjork 2015).

Traditionally, learning processes have been evaluated by measuring how the learner performs on tasks during the learning process itself. However, in a way that may appear counter-intuitive at first, it has been shown that performance during training can be misleading as a tool to judge learning improvement and training effectiveness (R. A. Bjork 1994). In other words, a good performance during training does not necessarily correlate with long-term learning measured by performance at later times.

In fact, scientific evidence supports the hypothesis that learning can be enhanced if the learner faces certain difficulties applying the newly learned knowledge or skills (E. L. Bjork, R. Bjork, and Mcdaniel 2011). These difficulties, known as **desirable difficulties**, are obstacles that make tasks during training harder for the learner and can negatively affect immediate performance, but are actually beneficial for learning and can improve the learner's performance in the future. The concept of desirable difficulties thus stems from the idea that long-term learning is fundamentally different from current performance. In this framework, to maximize long-term learning students should make things hard on themselves e.g. by i) using retrieval practice (testing or generating solutions), ii) having reduced feedback, iii) spacing practice sessions in time, and iv) interleaving different topics of study (Clark and Robert A Bjork 2014; Metcalfe 2011).

Following the framework of desirable difficulties and the idea that long-term learning does not necessarily correlate with performance during training, Kapur 2016 divides learning events into four classes:

- productive successes
- unproductive successes

- productive failures
- unproductive failures

In this classification, success or failure in a learning event makes reference to the learner's **performance**, which in this context is a measure of the **short-term** effectiveness of the learner in solving tasks. On the other hand, productive or unproductive is a reference to the effects on **learning**, a **long-term** measure of skills acquired. In particular, a productive failure happens when the learner faces a problem that requires new knowledge and, even when failing to find an optimal solution, long-term learning is favored. According to Kapur's characterization (Kapur 2016), for a productive failure to occur the task at hand should be solvable in multiple ways and challenging but not so difficult that the learner can loose motivation. Also, solving the task should involve making use of prior knowledge that the learner has, and instruction by a teacher or expert should follow the learner's attempt to solve the task.

Failing can then be beneficial to learning. But tasks that prompt failure of some kind are perceived as harder than tasks that learners complete without errors. However, as Clark and Robert A Bjork 2014 conclude in their review on the relevant literature on the introductions of difficulty and errors in the learning process, the offer of the desirable difficulties framework is the introduction of difficulty and errors to promote durable learning, instead of just increasing current performance.

In the context of learning in games, a study by Lomas et al. 2013 looked at engagement and performance in players of a single-player video game. Each player was randomly assigned a difficulty and they could keep playing for as long as they wanted. Similar to chess ratings, challenge was measured as the inverse of the chance of success in a game. They found that players tended to engage more - measured in time spent playing - as the game was easier. That is, challenge didn't promote engagement at all. However, they also found that players' learning curves improved faster when the game was harder, meaning that challenge was a positive factor in learning.

It's very plausible to see how the concept of desirable difficulties would apply in the realm on tournament chess: playing against harder opponents in challenging tournaments may result in poor immediate performance reflected in the loss of games and rating points. In the longer term, however, the player's live analysis during the game of the challenging positions and styles faced in those challenging tournament games provide an opportunity for training with unparalleled motivation. Other added benefits include the post-game analysis of the challenging games as well as getting more acquainted with the higher rated players and the environment. Finally, tournament games are very much like tests, which are highly recommended for learning: each time the opponent makes a move the player must find a correct answer, and do it without any help. Challenging games can be viewed, then, as tests in which the player must practice the best of it's abilities. Unchallenging games, would be, on the other hand, mere repetition of what is already known. This all compounds to make participation in tournaments against better players an experience that, though uncomfortable and possibly even unpleasant in the moment, could give a player a learning advantage against others in the future.

Of course, there may be opponents that are just too strong for the player, so that less benefit can be extracted from them. The players must be able to guide their search for answers on some knowledge they already have. If the problems that the player must solve in those games are so hard that the player can't even begin to understand, lack of connection to previous knowledge and loss of motivation would become detrimental to learning. The games that players loose to opponents that are too challenging, failing to discover correct answers by themselves, would be, in the terms described by Kapur 2016, unproductive failures.

1.2 Chess: Rating System and Studies of Performance Differences by Sex

In the world of competitive chess, the World Chess Federation, FIDE, manages a rating system that assigns a number to each player - a player's rating - that aims to measure the player's ability to play chess (FIDE 2021a). A player's rating is susceptible to change every time the player plays a game with a rated opponent. The change in rating after a game is a function of the difference in rating between the two players. The system implemented by FIDE, the Elo rating system, calculates the change to player A's rating, ΔR_A, after playing a game against player B and obtaining a result S_A –which can be 1 for a win, 0 for a loss or $\frac{1}{2}$ for a draw. First, an expected outcome for player A's result in the game is computed based on both player's previous ratings R_A and R_B as follows:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{1.1}$$

Then we can compute the change in player A's rating as

$$\Delta R_A = K_A \left(S_A - E_A\right) \tag{1.2}$$

Where K_A is the development coefficient of player A and is equal to 40 for a new player with less than 30 games, 20 for a player that has never reached a rating of 2400, and 10 for a player that has reached a rating of 2400 or higher at least once. In practice, this means that the rating of a higher rated player is less affected by the outcome of any individual game.

FIDE also awards players with titles when their rating has reached the following thresholds (FIDE 2021b):

- GM, Grand Master. 2300 rating points.
- IM, International Master. 2200 rating points.
- FM, FIDE Master. 2100 rating points.
- CM, Candidate Master. 2000 rating points.

Besides these titles, to which every player can aspire, FIDE has a special class of titles reserved only for players registered as women. The thresholds for these titles are:

- WGM, Woman Grand Master. 2100 rating points.
- WIM, Woman International Master. 2000 rating points.

- WFM, Woman FIDE Master. 1900 rating points.
- WCM, Woman Candidate Master. 1800 rating points.

It is important to note that even though there exists a class of titles awarded only to women, the rating system is one and the same for all players, regardless of their registered sex. This makes comparison of players' ratings across sexes possible. Also important for this point is the fact that players of both sexes interact in play in competitive tournaments at almost every level of the game. This is in contrast with most competitive activities where there exist separate exclusive classes or leagues for men and women. However, it is also important to note that in chess there do exist many women-only tournaments in every age category and in almost every rating level. This conforms an interesting competitive environment to analyse in which players can mix with each other, in terms of their registered sex, but at the same time, in the case of women, can choose to segregate themselves.

This unified system for all players has made chess, and the community of players of the game, historically used by scholars as a testing ground for hypotheses in psychological and cognitive science. Among the reasons to use chess to study cognitive processes, the most relevant are the fact that ability is objectively quantifiable thanks to a clear and common rating system, and that there is availability of data spanning many years (Howard 2005; Vaci and Bilalić 2017).

According to Vaci and Bilalić 2017, there have been two main lines in which chess has been used as a vehicle in psychological studies: first, to determine how experts differ from novices and the factors that contribute to that difference. For instance, studying the different ways in which experts and novices practice and correlating these difference with the corresponding ratings can provide insight into what the best strategy is when learning in an intellectual field like chess.

Notably, deliberate practice, a theory of how expertise is developed, has been studied in the realm of chess. Charness, Tuffiash, et al. 2005 argue that deliberate practice is most present in the self-study part of chess practice and that this accounts for the greater correlation of this type of activity with rating performance in tournaments. They also note that tournament play is a type of practice with unique characteristics, such as the variability of the opponent's skill – and the fact that it's beyond the practitioner's control - the need to efficiently manage time, and the challenge of concentrating in a distracting environment. These findings are coherent with our understanding of tournament play as tests in the player's learning process.

The second way in which chess has been used in psychology and cognitive science is in the study of relationships between variables of interest and the performance of players. In this case the variables of interest include age (Fair 2007; Roring and Charness 2007), practice (Charness, Tuffiash, et al. 2005; F. Gobet and Campitelli 2007) and sex (Bilalić, Smallbone, et al. 2009; Blanch, Aluja, and Cornadó 2015; Charness and Gerchak 1996; Howard 2005, 2014; Knapp 2010; Maass, D'ettole, and Cadinu 2008). Other examples are the study of the effects of belonging to a minority and the Einstellung effect – the decrease in expert's performance when the first solution found is not the optimal (Vaci and Bilalić 2017). Also, Bilalić, McLeod, and Fernand Gobet 2007 compared personality traits of children and their participation in chess and propose that agreeableness, a personality trait for which girls score higher on average than boys, may be an factor in determining the differences in participation rates in chess.

The literature that examines the chess competition environment to study sex differences works in two main complementary efforts: the description and quantification of those differences, and the proposal of plausible interpretations and explanations. Howard 2005 analyzed chess ratings to quantify sex differences in performance, see if differences had been diminishing with time and study possible explanations that could be proposed by looking at the data. To analyze current sex differences, he studied the whole population of players in the January 2004 rating list. To observe the evolution of differences, he analyzed rating lists of the previous 50 years.

The first thing to notice in this study, is the enormous difference in participation rates. In the January 2004 rating list, for instance, there were 50 450 players, of which only 3 646 were women. In the study of that rating list, Howard found that male advantage in the ratings was clear: men's mean rating was one standard deviation higher than women's when all ranked players were taken into account. In addition, it was found that male's ratings distribution was more variable than women's. Also, when only the best n players of each sex were considered, with n being the number of women in the list, to compare the same number of players of each sex, men's mean rating was over two standard deviations higher than women's. Comparing mean ratings of the best n players in each year showed there had been no decrease in the rating gap over the past three decades.

Additionally, Howard's research analyzes the career patterns of every male and female that entered the ratings list from July 1985 to July 1989 and found that only 38% of women were still active past 1999 vs. 67% of males. Also, men played many more games during their career and this difference was larger in the top 100 of the analyzed players, where males in average played three times more games than females. For Howard, the differences in chess careers of males and females may have multiple interpretations such as differences in obsession levels with the game, differences in natural talent or differences in life choices –chess may only be a passing interest for most women. Howard concludes that the results are consistent with the view that the sex differences at high achievement levels are partly due to ability differences, as opposed to being caused by purely social factors such as the glass ceiling - obstacles that hinder the advancement of women's careers - or the lack of female role models.

Different approaches have been attempted to try to explain the gap between men and women seen in chess ratings. Usual ideas for explanations include: innate biological differences, cultural differences and social obstacles and stereotypes. For instance, Maass, D'ettole, and Cadinu 2008 showed that a common social stereotype in chess, that men are supposed to be better than women, does have an influence on how good a player performs once they know the sex of their opponent.

On the other hand, Chabris and Glickman 2006 studied the rating lists of the United States Chess Federation (USCF) for the previous 13 years. They included in their analysis the number of games played and the age of the players. Their analysis found large differences in mean rating in favor of males but no evidence of higher variation in the male population. However, when they analyzed the rating evolution of male-female pairs during 3 years arranged on the basis of equal initial rating, they found that that the rating difference within the pairs wasn't significantly different to zero. Also, when they analyzed the ratings of players 6 - 12 years old - to test the hypothesis of different initial abilities between male and female - in regions where girls were at least 50% of the playing population, they found no significant sex difference in the mean ratings. In the view of the authors, their results support the idea that the observed sex differences in ratings can be attributed to different participation rates, an idea commonly known as the "participation-rate hypothesis".

Under this view, even if the ability distributions of both populations – men and women – had the same means and the same standard distributions, almost all of the difference in representation at the top of the ratings would be explained by the difference in participation rates (Bilalić, Smallbone, et al. 2009; Charness and Gerchak 1996). Practically speaking, this explanation says that there are many more men at the top of the chess rankings because there are many more men playing the game. Therefore, the difference in participation rates would have to be taken into account before attempting any other explanation. Following this principle, Bilalić, Smallbone, et al. 2009 analyzed the rating list of the German chess federation for 2007, with 120 399 players, and concluded that the difference in participation rates explains 96% of the observed rating difference.

However, the suggestion that the difference can be explained, for the most part, by statistical reasons, has raised objections. Knapp 2010 argues that the model of Bilalić, Smallbone, et al. 2009 is inadequate on the basis that it predicts ratings for the top players so high that they are in conflict with the ratings found in reality and proposes a hypergeometric distribution for the ratings instead of the normal distribution assumed by Bilalic *et al.* He then argues that in this case the difference in participation rates can only explain up to 71% of the difference in performance at the top. Also, Howard 2014 criticised their work arguing they assume that players from both sexes are drawn from the same part of the respective underlying distribution of each sex.

Blanch, Aluja, and Cornadó 2015 also analysed ratings and concluded that differences in participation rates are insufficient to explain the difference in ratings by themselves. Their analysis points to other factors – such as age and practice – as variables that also influence the rating differences.

A further study by Howard 2014 analyzed the ratings taking into account number of games played and players' performance limit - the point at which they don't improve anymore. He found that a greater proportion of males with more than 750 games achieve a Grand Master title than the proportion of women with more than 750 games that obtain it. Also, the proportion of women in top rating positions does not improve in countries where the female participation rate is much higher. He concluded that social factors are not the only cause for the observed data. Howard concedes that deliberate practice was not taken into account, and so a criticism on his research could be that the observed rating differences may be due, for instance, to the fact that male players simply study more. To this, his response is that his work analyzed ratings at player's performance limits and there is no evidence supporting that more practice of any kind could improve their performance.

1.3 Agent-based Modeling

In the study of complex systems with many individual agents and interactions between them, it is often the case that a relatively simple model can be formulated describing how each of the individual agents interacts with one another. An agent-based model, or ABM, is a formulation of this type in which each individual, called an agent, has a set of properties with values of it's own, and interacts with other agents according to an established set of local rules (Railsback and Grimm 2011).

A simulation of an agent-based model is a computer program than implements the formulation of an agent-based model. It runs an algorithm that defines the agents and properties, and simulates the evolution of the system by applying the interaction rules described in the model. This allows the observation of changes in the system, or a part of it, and the emergence of properties on a system-wide scale. As a computational method, ABMs are generally understood to include the computer simulation that implements the model. The program for this simulation can be coded on general purpose coding language, or on any of the languages and software specifically oriented to building ABMs (Abar et al. 2017).

The ABM approach can be very useful when there is interest in studying the general properties of a complex system but there are no known analytical solutions describing the behavior of these properties. Well designed models and simulations can be useful in various ways, such as making predictions, finding counter-intuitive outcomes or supporting an argument against previously accepted theories, among others (Epstein 2008).

The building and use of agent-based models in the study of social systems

Published research on Agent-Based Models (ABMs) in the last few years, as a field, has two main features: i) an expanding array of sciences in which ABMs are being applied to investigate research questions, and ii) efforts to propose general frameworks and guidelines for the building of ABMs and the interpretation and validation of the simulated results.

Examples of publications with at least 50 citations in the last 4 years include topics as varied as cancer cell growth (Z. Wang et al. 2015), anthropology (Dyble et al. 2015), energy technology adoption (Rai and Robinson 2015), emotional dynamics (Bosse et al. 2015), economic dynamics (Assenza, Gatti, and Grazzini 2015), economic crisis resolutions (Klimek et al. 2015), decision making and safety (H. Wang et al. 2016), and gender inequality (Grow and Van Bavel 2015).

The proliferation of fields of application of ABMs, and the specific problems in which they are being applied, can be better understood with the consideration that ABMs are representations in which a system is modeled by modeling the behavior of its individuals. In this manner, the system's behavior is not the product of a system-wide program, but the consequence of the behavior of the individuals. The collective properties observed in the system are then emergent phenomena. This characteristic makes ABMs attractive to study social phenomena. Indeed, in the social sciences, ABMs are part of a broader field that has come to be known as Computational Social Science which, in the view of Cioffi-Revilla 2014, has two main aims: to understand the social universe, and to improve the world in which we live.

As part of the efforts to consolidate the broader application of computational techniques in the social sciences, in a paper titled "*Manifesto of Computational Social Science*" (Conte et al. 2012) published in 2012, key authors of research on simulation in the social sciences identified the field's focus on social phenomena that emerges from individual's behavior such as segregation, cooperation, reciprocity, social norms or institutions.

Macy and Willer 2002 recount the history of social simulation in the past century. In their account, ABMs are a third wave of social simulation techniques that start coinciding with the advent of the personal computer in the 1980s. These new models were different from the ones in the previous two waves of social computing in that they now modeled individual characteristics and behaviours, as opposed to the macro characteristics modeled in the first wave, and in that the individuals were able to interact with one another and adapt, contrary to the micro but isolated behaviour modeled in the second wave (ibid.).

ABMs are used in a wide variety of fields (Macal 2016). But, not only that, the building of one single model can incorporate knowledge from a wide variety of fields itself. One of the first ABMs, Axelrod's prisoner's dilemma tournament, for instance, included strategies submitted by professors of economics, political science, psychology, sociology and mathematics (Axelrod 2006).

ABMs, by their bottom-up approach to the representation a system, are seen now as a very useful tool to study collective consequences of individual behavior because they link individual actions with their emergent collective outcomes (Bruch and Atwell 2015). ABMs are perceived by a growing number of researchers as a shift in the paradigm of social science and it's popularity continues to rise (Macal 2016).

One advantage of the ABM approach over other simulation techniques is the agent perspective that is taken when building the model (ibid.). Another advantage, and maybe one that explains it's growing use, is that ABMs are particularly useful when the system being modeled requires the representation of heterogeneous actors with interaction dynamics that are difficult to control with other mathematical representations (Cioffi-Revilla 2014).

Elements of the ODD protocol	
Overview	1. Purpose
	2. Entities, state variables and scales
	3. Process overview and scheduling
Design concepts	4. Design concepts
Details	5. Initialization
	6. Input data
	7. Submodels

Table 1.1: Elements of an agent-based model description following the ODD protocol (Grimm, Berger, DeAngelis, et al. 2010)

Despite the growing predominance of ABMs, common methods and accepted general practices are not established (Macal 2016). This statement is a common theme in the agent-based modeling literature and is the reason various authors have proposed generalized methods for building, studying and validating ABMs. The importance of this issue can be summarized by Grimm and Railsback's statement that it is hard to produce science with ABMs without a systematic approach to their formulation and description (Railsback and Grimm 2011).

The ODD protocol - an acronym for Overview, Design concepts and Details - has been proposed (Grimm, Berger, Bastiansen, et al. 2006) and updated (Grimm, Berger, DeAngelis, et al. 2010) with the objective of making a systematic description of ABMs. The purpose is to make the models understandable and not difficult to duplicate. This is expected to help in making research that use ABMs reproducible (Grimm, Berger, Bastiansen, et al. 2006). Table 1.1 lists the elements that a description of an ABM should make explicit when following the ODD protocol (Grimm, Berger, DeAngelis, et al. 2010).

The elements in the ODD protocol describe different aspects of the model. Since the protocol is a standard, they should be listed in the same order as they appear in the standard formulation, that is, in the order given by Table 1.1 (ibid.). Each element should describe the following aspects of the ABM(ibid.).

1. **Purpose**. The purpose or objective(s) of the model

2. **Entities, state variables and scales**. The entities are units or actors of the model. The variables that describe each entity's state and that makes it distinguishable from others. Usually, the kinds of entities used are: Agents, spacial units, environment, and collectives.

3. **Process overview and scheduling**. A detailed description of the processes of the model in terms of the actions and the entities that perform those actions. These will be the submodels in the *submodels* section of the ODD protocol. Also, a description of how time is handled in the model. The inclusion of pseudo-code in the description is recommended.

4. **Design concepts**. Concepts that make the reader aware of the conscious decisions made when designing the model. Models usually include these, but if one or more elements are not part of a particular model then it may be excluded from its description.

    - Basic principles
    - Emergence
    - Adaptation
    - Objectives
    - Learning
    - Prediction
    - Sensing
    - Interaction
    - Stochasticity
    - Collectives
    - Observation

5. **Initialization**. The initial state and conditions used when running the model. This is crucial to allow replication of the experiments run with the model.

6. **Input data**. A description of the external sources or input data that the model uses, or the explicit statement that it doesn't.

7. **Submodels**. A description of the processes listed in the *Processes overview and scheduling* section including parameters and dimensions and reference values of the processes.

Chapter 2

A Study of Segregation and Performance in FIDE-rated Games

In this chapter we present our findings after performing a statistical analysis of data we collected of a group of players and the games they played during a 10-year period. This aims to contribute to the two usual goals in the study of sex differences in chess: describing the differences and presenting possible explanations for them. To this end, we describe differences found associated with sex and, as part of contributing to an explanation, we present and discuss the relationship found between the challenge players face in competition and their improvement in rating.

2.1 Sample composition

The games in the study correspond to a 10-year period from January 2010 to December 2019. In this period, all registered game pairings for a group of 1463 players were analyzed. We selected players that started the 10-year period with a rating lower than 2400 FIDE rating points and ended 2019 with a rating above 2350 FIDE rating points, i.e. they improved or, in the worst case, nearly maintained their rating.

These players were classified into three groups according to their registered sex and their rating in Dec 2019, the end of the period, as follows: players of either sex that finished above 2550 were assigned to the **Top improvers** group, women that finished with a rating below 2500 were assigned to the **F improvers** group, and men who finished with a rating below 2500 were assigned to the **M improvers** group. The composition of the groups is shown in Table 2.1. Players that finished 2019 with a rating between 2500 and 2550 were excluded in order to introduce a gap between the Top improvers and the other groups, and thus explore factors that may contribute to a qualitative difference in performance.

We wanted to explore if challenge differences may be associated with women not reaching the top positions in the rankings. In this sense, the threshold rating of 2550 for the Top improvers group was chosen because there are very few women above this rating: out of 511 players above 2550 in the January-2020 FIDE ratings, only 9 are women.

	Male	Female	Total
Finished above 2550 (Top improvers)	69	1	70
Did not finish above 2500	1340 (M improvers)	53 (F improvers)	1393
Total	1409	54	1463

Table 2.1: Composition of groups of players analyzed from 2010 to the end of 2019

2.2 Findings

Figure 2.1 shows the rating evolution by group for every player in the study during the 10-year period. We can see the wide rating range at the start of the period for the three groups, but final ratings are constrained by the defining condition on each group, so Top improvers are clearly better in the end. We can also observe how for almost all players that make really high improvements, most of the rating climb is done in the first five years of the 10-year period.

Figure 2.1: FIDE rating every semester from January 2010 through December 2019 by group.

We found that women in the sample play more against other women than against men (see Figure 2.2). In the 10-year period, women played 61.5% of their games against other women. For comparison, men in the sample, who could only play in open tournaments, played 4.78% of their games against women. This low percentage makes sense given that women are only a small fraction of the population, so the chance of playing against a woman in an open tournament is also small. We can infer from this that the fact that most women in the study played the majority of their games against other women means that women participate more in women-only tournaments than in open tournaments.

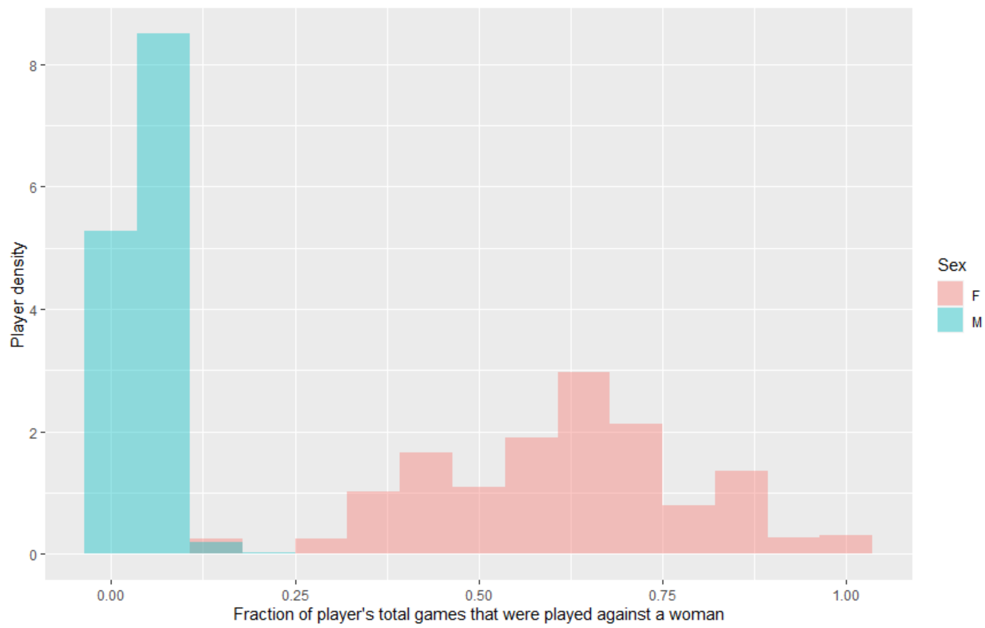

Figure 2.2: Distribution of the fraction of their games that players played against a woman, by sex.

We are interested in examining variables related to segregation, activity, and challenge, and their relation with rating improvement. In Table 2.2 we can see the three groups studied and different measures of activity, challenge and segregation. Only years in which a player started with a rating between 2400 and 2500 were counted there. This rating range is of interest to our work because it corresponds to ratings just under the threshold that the Top group overcomes, but the other groups don't. The variable **Games played per year** is a measure of how active the players of a group were on average, regardless of the opponents they played. **Games vs 2550+ rating**, in absolute number per year and as a percentage of the number of games played, are measures of how much the players competed against the best players in the world. **Games with 100+ challenge**, in absolute number per year and as percentage of games played, are, on the other hand, measures of how much the players faced challenging games against players that are admittedly better (less than 0.37 expected outcome). Last, **Games vs. women** as a percentage of the number of games played is useful to measure the average level of segregation (in the case of F improvers) but also, looking at the value for Top and M improvers, this measure shows the percentage of games against women by players that only play in open tournaments. The data show that players that ended in the Top group played more games than players in the other groups, and also that women played more games on average than the men in the M improvers group. We see that the top players played more against more challenging opponents, both in absolute number of games and as a fraction of their total games played. Women played less challenging games than the top players. Of the three groups, women in the F improvers group in the 2400-2500 range played the smallest percentage of their games against challenging opponents.

The findings show that greater activity with greater challenge was associated with greater performance improvement (Top improvers played more and had more challenging games), while at the same time, greater activity without greater challenge did not lead to better performance improvement (F improvers played more games, but less challenging ones than the M impovers, ending in the same rating range).

	Games played per year	Games vs. 2550+ rating	Games vs. 2550+ (% of total)	Games with 100+ challenge	Games with 100+ challenge (% of total)	Games vs. women (% of total)
F improvers	65.3	5.80	8.89%	7.24	11.1%	70.1%
M improvers	43.6	4.84	11.1%	5.81	13.3%	4.64%
Top	80.4	16.0	19.9%	13.1	16.2%	5.91

Table 2.2: Measures of activity, challenge and segregation for each group. Values are averages per year of all players in each group.

We further analysed the relation between challenge and performance improvement by confronting different measures of challenge with rating improvement during a year i.e. $R_{\text{end of year}} - R_{\text{start of year}}$. In Figure 2.3 we can see the relationship between rating improvement and number of games vs. 200+ challenger for players with 40 or more games. For all three groups - Top, M and F improvers - there was a positive correlation between improvement in the year and the number of games against players 200 or more points better (0.64, 0.48 and 0.39 Pearson coef., respectively). On the other hand, we observed a weak-moderate negative correlation (-0.54, -0.44 and -0.34 Pearson coef., respectively) between the fraction of low-challenge matches and rating improvement (Figure **??**). These two facts combined give support to the applicability of the idea of desirable difficulties in the chess competitive environment: that improving in chess is helped by participating in challenging games.

Figure 2.4 disaggregates Figure 2.3 for each year in the period studied. We can see that the overall relationship observed in Figure 2.3 holds for the first half of the period, but cannot be observed in the later years. This is due to the fact that there is little rating change found in the second half of the 10-year period, which makes sense, as more players may already be near their peak rating and they show lower amounts of rating improvement. Also, specially in the case of Top improvers, most players show low improvement and low challenge, as there are less players that are better than them to compete with. This last point also puts into context the overall challenge of the Top group, showing that most of the time in the study they were already at their best, with no players that are hundreds of rating points above.

By focusing in the earlier period (years 1 through 5), when players are still improving their rating in all groups and experiencing different levels of challenge, we can better explore the relationship between the two. To compare the possible effects of facing different levels of challenge, several variables measuring games played within specific challenge ranges were evaluated. The Pearson correlation coefficients of these variables measured for each player by year (years 1 through 5) with the rating improvement in that year are shown in Table 2.3.

In period of years 1 through 5, the challenge range variable with the highest correlation with improvement was the *number of games played with a challenge between 250 and 350 points* (0.45 Pearson). On the other hand, in the same 5-year period, the *fraction of games played with less than 50 challenge* showed a moderate negative correlation with rating im-

Figure 2.3: Rating gained in one year vs. number of games played in that year with a challenge of 200 points or more. Each point corresponds to a year of an individual player. For each player, only the years with 40 or more games are shown.

	Correlation coeff. with rating improvement in year			
	All groups	Top improvers	M improvers	F improvers
Games played in the year	0.16	0.06	0.15	0.09
Number of games with less than 50 challenge	-0.12	-0.22	-0.13	-0.11
Number of games 50 - 150 challenge	0.20	0.02	0.22	-0.04
Number of games 100 - 200 challenge	0.30	0.31	0.29	0.14
Number of games 150 - 250 challenge	0.39	0.47	0.37	0.33
Number of games 200 - 300 challenge	0.43	0.53	0.42	0.37
Number of games 250 - 350 challenge	0.45	0.54	0.44	0.38
Number of games 300 - 400 challenge	0.44	0.49	0.44	0.41
Number of games 350 - 450 challenge	0.43	0.46	0.43	0.36
Ratio against less than 50 challenge	-0.45	-0.49	-0.45	-0.35
Ratio against 50 - 150 challenge	0.14	-0.03	0.17	-0.09
Ratio against 100 - 200 challenge	0.25	0.30	0.25	0.13
Ratio against 150 - 250 challenge	0.04	0.48	0.33	0.37
Ratio against 200 - 300 challenge	0.39	0.52	0.38	0.38
Ratio against 250- 350 challenge	0.40	0.53	0.40	0.38
Ratio against 300 - 400 challenge	0.40	0.46	0.41	0.42
Ratio against 350 - 450 challenge	0.38	0.42	0.39	0.37

Table 2.3: Pearson correlation coefficient of different measures of challenge in a year with the rating improvement in the same year.

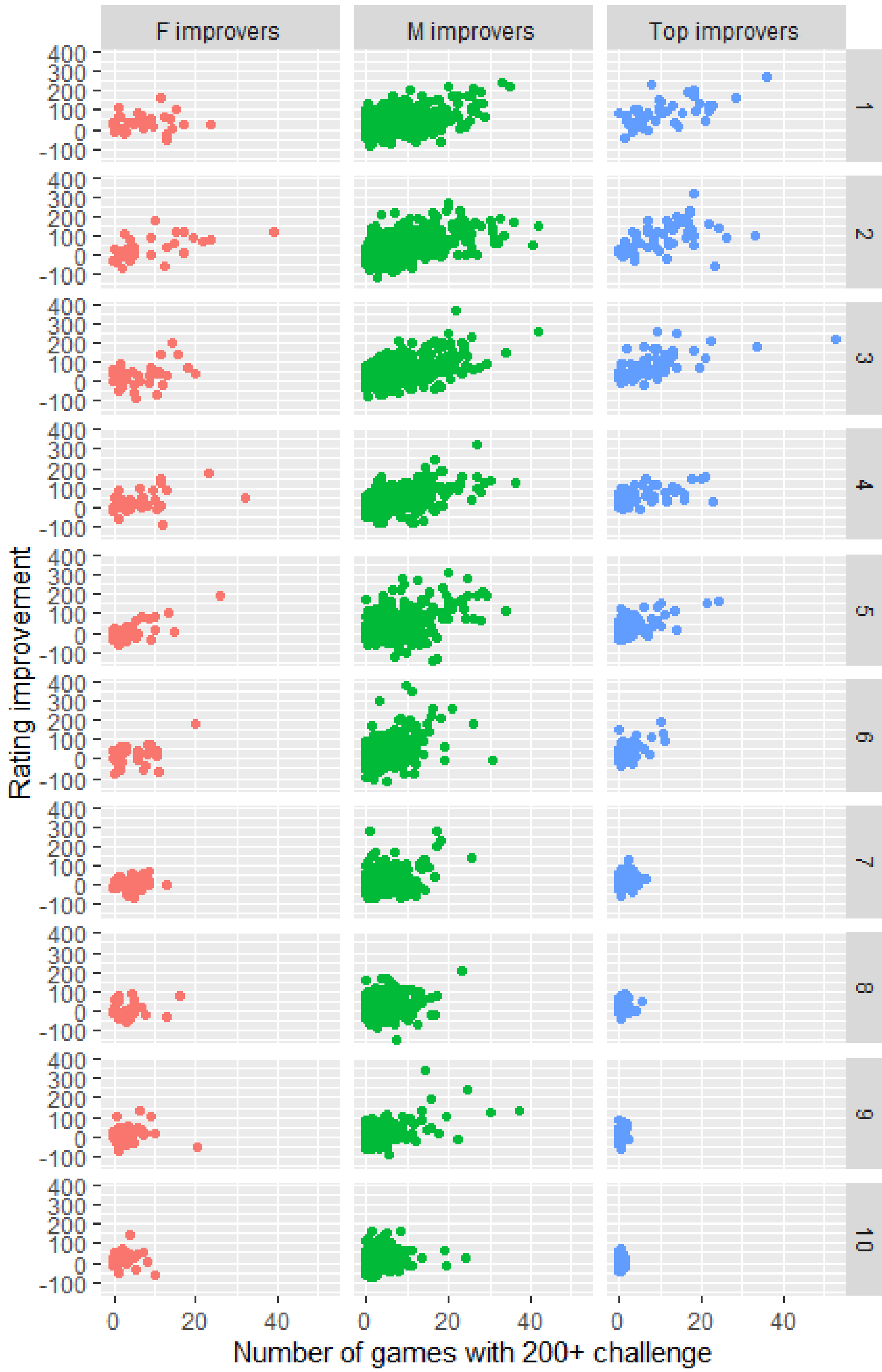

Figure 2.4: Rating gained in one year vs. number of games played in that year with a challenge of less than 50 points, by group and year. Each of the 10 years in the period is a row of panels. For each player, only the years with 40 or more games are shown.

provement (-0.45 Pearson). It's also worth noting that just playing many games was not by itself associated with becoming a better player: the correlation between rating improvement and the number of games played was found to be very weak (0.16 Pearson). This strengthens the idea that it's not just the number of games played that's important, but the number of those games that are actually challenging. Overall, as well as for each group, here we see again rating improvement having positive correlation with challenge and negative correlation with the ratio of games that are not challenging.

It's worth noting also that while the negative correlation of improvement with the ratio of games against less than 50 challenge is moderate, the negative correlation with the number of games with that same low challenge is very weak. It seems then, that playing many games with low challenge is not necessarily bad for improvement, as long as those games are not a large fraction of the total games played.

Chapter 3

The Agent-based Model

In this chapter we present an Agent Base Model (ABM) to address the question of how segregation could affect performance improvement. Or goal is to build an ABM representation of a system of competing chess players consistent with the theory of desirable difficulties and run computer simulations of it to to analyse the relationship between segregation and sex differences in performance.

In the interest of standardization, the presentation of the model will be guided by the ODD protocol.

3.1 Purpose

The purpose of the ABM is to represent a system of competing male and female chess players that incorporates the concept of desirable difficulties - that some level of challenge is beneficial to learning and long-term improvement. By this implementation, the goal is to inspect any emerging connections between the preference for same-sex segregation in women and the comparative performance of the top women players.

3.2 Entities, state variables and scales

Each agent in the ABM represents a player and has the properties listed in Table 3.1, with static properties being constant throughout a simulation run and dynamic values being modified.

Players are grouped into tournaments, basically lists of potential opponents during each step in the simulation. The lists of properties each tournament has is shown in Table 3.2. Tournaments are uniquely identified by their type/number combination. Tournaments are destroyed after they are finished and new ones are created.

3.3 Process overview

Players in the ABM compete against each other by playing games between pairs of opponents. Game pairings are assigned by first assigning each player to a tournament. Once all players

Sex	Static. Either Female or Male, representing players' registration as women or men in the players list.
Group	Static. Set as *Men* for all Male players, and either *Test* or *Segregated* for Female players.
Rating	Dynamic. Positive integer representing how good a player is. Can be viewed as the FIDE rating actual players have.
Rating modification factor (K)	Dynamic. A positive integer used to make rating less volatile for higher rated players. It's equivalent to the K modification factor in FIDE's rating calculations.
Segregation preference	Static. Only relevant for female players. Gives the probability that the player will choose a women-only tournament each time tournaments are assigned.
Learning	Dynamic. Real, positive number quantifying the learning gained by the player in previous games.
Past learning benefits	Dynamic. A list of the previous learning benefits obtained by the player.
Current tournament type	Dynamic. The type of tournament in which the player is currently participating.
Current tournament	Dynamic. The ID of the tournament in which the player is currently participating.
Wins in current tournament	Dynamic. The number of games won in the current tournament

Table 3.1: Properties of players

Tournament type	Static. Either *Open*, in which players of both sexes can participate, or *women-only*, in which only female players can participate.
Tournament ID	Static. A positive integer identifier for the tournament that is unique within its type.
Tournament size	Static. The number of players currently in the tournament.
Tournament rounds	Static. The number of rounds that must be played to complete the tournament.
Current round	Dynamic. The number of rounds that have been played in the tournament. The tournament will be finished after a pre-determined number of rounds.

Table 3.2: Properties of tournaments

are assigned to a tournament they all play a fixed number of games against opponents within their tournaments. After this, new tournaments are assigned to all players. At any given time, each player is assigned to exactly one tournament. There are two kinds of tournaments:

- Open tournaments (Male and female players can play)
- Women-only tournaments (only female players can play)

Figure 3.1 shows a flow chart of the general processes of the ABM. All the logic involved in the application of the LFC model, the calculation of game outcomes and the recalculation of ratings after a game takes place in the *Play game* sub-process.

3.4 Design concepts

Basic principles

A core principle in the ABM design is the concept of Desirable Difficulties, i.e., the idea that challenging tasks, that are neither too easy, nor excessively difficult, benefit mid and long-term improvement (E. L. Bjork, R. Bjork, and Mcdaniel 2011). The application of this idea in the ABM is central for the conceived mechanism through which it is hypothesised that group differences in performance may emerge between male elite players and female elite players who segregate.

The ABM allows us to study the effects of segregation by splitting the population of female players into two groups, *Test* and *Segregated*, and assigning each group a different level of segregation. Each of the two female groups will have one half of the total female agent population. One of the two female groups, the *Segregated* group, will always play in women-only tournaments. The other female group, the *Test* group, can have a different segregation level, or strategy of segregation. The *Test* group, for instance, could not segregate themselves at all and play all their games in open tournaments, or segregate themselves completely and always compete on women-only tournaments along with the *Segregated group*, or adopt any intermediate strategy between these two. The precise strategy of segregation of the *Test* group is determined by the *Segregation preference* parameter P_{seg} of the model, which is a real value in the range $[0, 1]$ representing the probability that a player of the *Test* group chooses to play in a women-only tournament each time tournaments are assigned in the ABM. The value of P_{seg} is kept constant during a single simulation run of the ABM and can be changed from run to run to explore the effects of different levels of segregation.

Emergence

The two female groups are initially set from the same rating distribution. The only difference between these two groups is the preference for segregation, that is, how much they play against other women. Emerging group differences in rating, such as the rating of the top players in each group, will then be associated with the difference in segregation.

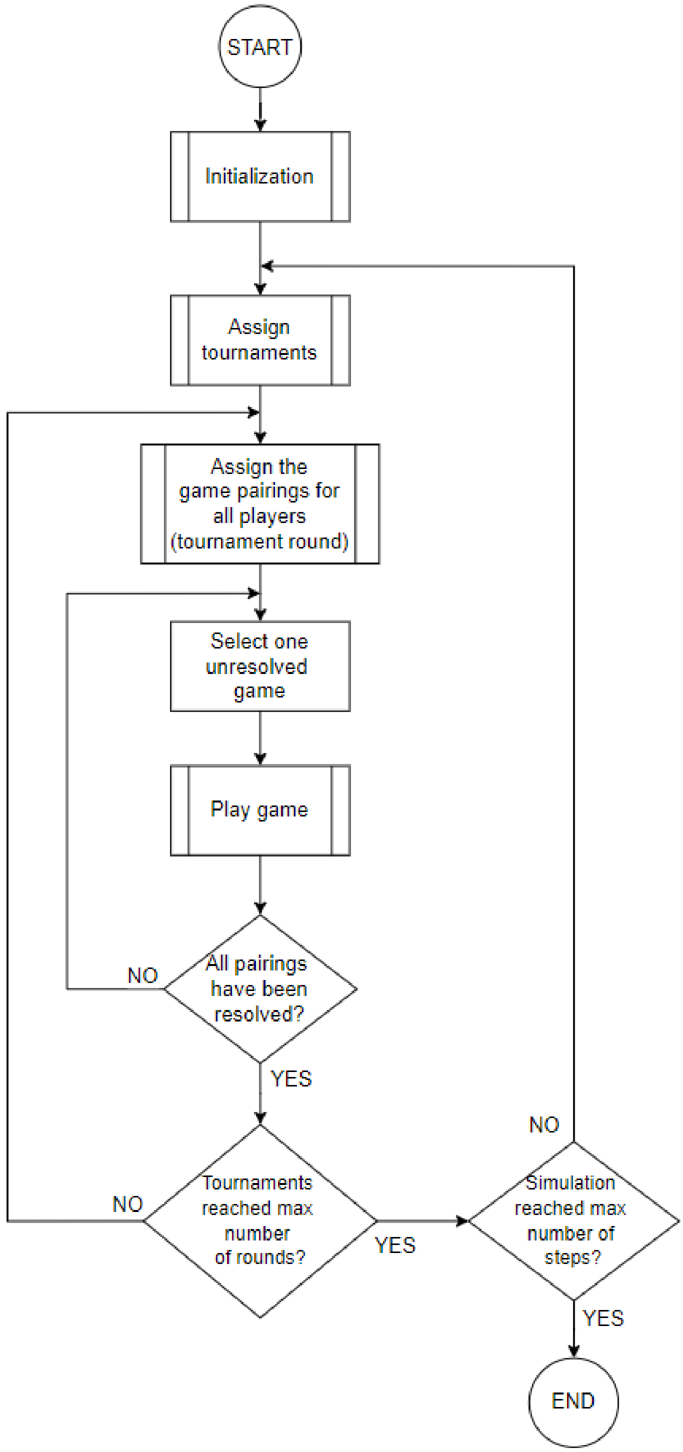

Figure 3.1: General flow chart of the ABM's processes.

Learning

Every time a game is played, the players involved acquire learning. This is independent of the rating change that follows the result. It depends only on the skill difference between the players. This learning is compounded with learning from previous games and will influence the players probability of winning the games that follow.

Interaction

Players interaction with each other are limited to playing games. In this interactions, both players involved update their learning and their rating. In the case of the winner of the game, the number of wins in the current tournament of that player is also updated.

Stochasticity

Two sub-processes are modeled to include randomness: self-segregation choice by players in the test group and deciding the outcome of the games.

Every time new tournaments are created and each player is entered into one of them, female players in the *Test* group choose what type of tournament to enter, open or women-only. This choice is modeled as a random value with probability P_{seg} of choosing a women-only tournament, where P_{seg} is a global parameter common to the entire *Test* group.

When two players play a game, randomness also plays a part: the probability that player A wins a game against opponent B is a function of the sum of the difference between the players' skill levels, which are represented by the sum between rating and previous learning. The ABM does not involve a simulation of a game with the rules of chess, it simulates the result.

Collectives

Players belong to groups, *Men*, *Test* and *Segregated*. The group a player belongs to is the defining factor in determining the kind of tournament the player enters.

Once the type of tournament is defined, a player takes part in a specific tournament of that type according to his/er rating. The tournament an player is in determines the pool of his/her possible opponents in the next few games.

Observation

In each group, the ratings of the top 20 players are observed and are the values on which the group performances are compared.

3.5 Initialization

Initial values of the agent's properties require two rating lists and a parameter: a list of male players with their ratings, a list of female players with their ratings, and the fraction of the agent population that will be female P_w. The agent's properties are determined as follows:

- *Sex*: For each player, *sex* is assigned randomly with probability P_w of being female.
- *Rating*: Each player is assigned a *rating* sampled without replacement from the provided rating list of the same sex.
- K: Emulating FIDE's model, the K value is 20 for players below 2400 *rating* and 10 for players with *rating* 2400 or more.
- *Group*: each player's *group* is assigned according to the following rules:
  - All male players are assigned to the *Men* group
  - Half of women are assigned to the *Test* group
  - The other half of women are assigned to the *Segregated* group.

3.6 Input data

On initialization each player is assigned a starting rating. In this work we have used FIDE's complete rating list of January 2010 for both men and women as input lists from which we sample to assign initial ratings to the players in the ABM. This means that the rating distributions of both men and women in the ABM resemble the actual rating distributions of chess players at the beginning of 2010.

The ABM also requires some parameters as inputs. Table 3.3 shows the full list of the ABM parameters grouped in three categories according to which part of the model parametrization they are related to. These parameters relate to the mathematical model of learning from challenge which we will explain in detail in Section 3.7.

3.7 Submodels

3.7.1 A mathematical model of learning from challenge

The ABM needs to incorporate some mechanism that represents how players benefit from encountering challenge in the games they play, thus acquiring learning that can be useful in future games, a mathematical model of learning from challenge. If we start by taking into account the possibility of learning from challenging experiences, then a player's ability must be described not only by their performance in recent games but also by what they learned through playing in them, regardless of whether the games were won or lost. This learning may not yet have been reflected in their rating. A player may have just lost a game, loosing rating, but his or her ability may still have improved because of learning in that lost game. In this sense, there is an intrinsic ability not entirely represented by rating. We can then think of an *intrinsic rating* I representing a player's ability and define it as

$$I = R + L \tag{3.1}$$

where R is the official rating determined by game outcomes, and $L \geq 0$ is the part of the player's ability due to learning in previous games independent of their outcomes.

As a player faces new opponents, new learning is acquired from those games. Both their rating and their learning change with time as a consequence of playing games. In our model, learning, just like rating, only changes through playing games. Time in the model is a discrete variable that increases by one at the moment of playing a game. It can be viewed as a counter of the player's games. A player arrives at a game at time t with a rating R_t and learning L_t and leaves the game at time $t+1$ with rating R_{t+1} and learning L_{t+1}. After the game, a player's rating and learning will not be modified until the next game.

Population composition parameters	
Number of players	The total size of the population being simulated
Women's participation rate P_w	The fraction of the total player population that is female.
Segregation preference of the Test group P_{seg}	The probability that a female player in the test group chooses a women-only tournament.
Tournament parameters	
Standard tournament size	Number of players that each tournament should have. Because of the total number of players in each type of tournament (open and women-only), the size of the last assigned tournament of each type may vary.
Number of rounds	The number of games that each player plays in a tournament before it is completed. It is the same for all tournaments.
Learning model parameters	
Ideal challenge	A player gets the maximum learning benefit in a game when facing an opponent who is better by this number of rating points.
Maximum learning benefit	The maximum amount of learning benefit that can be obtained in a single game.
Benefit spread	A measure of how far from the ideal challenge can a player face while still obtaining significant learning benefit in a game.
Number of games in learning history	The number of future games in which the learning from any specific game still has influence.

Table 3.3: Parameters of the Agent-based model and their meaning

To apply the concept of desirable difficulties, the learning l obtained by a player in a game, that is, the difference between L_t and L_{t+1}, has to be related to the challenge that the player faced in that game. The intrinsic rating I, since it includes learning, is therefore a function of how much a player was challenged by their opponents. Challenge in a single game, i.e., a measure of how difficult it is for the player to overcome their opponent, is defined in our model as the difference in rating between the players involved.

$$c = R_{opponent} - R_{player}. \tag{3.2}$$

Any learning gained thanks to the challenge faced in that game would then be a function of that challenge:

$$l = f(c) \tag{3.3}$$

The model should reflect the idea that to learn better one should "*make things hard on oneself, in a good way*" (E. L. Bjork, R. Bjork, and Mcdaniel 2011), according to which tasks should neither be easy, nor excessively difficult, to have the most potential to improve learning. That is, in any game the learning benefit l as a function of the challenge c should peak at a certain value, some ideal rating difference, and go to zero as the difference in rating c becomes greater or smaller moving away from that ideal. A simple and smooth function that presents this behavior is e^{-x^2}. And so, a simple model for the learning benefit l associated with playing against an opponent who poses a challenge c in a game is (see Figure 3.2):

$$l = l_{max} e^{-\left(\frac{c-C_i}{S}\right)^2} \tag{3.4}$$

where l_{max} is the maximum learning benefit possible from a game, C_i is the ideal challenge from which a player would learn the most, and S is a measure of the spread of the l-vs-c curve indicating how quickly or how slowly learning decreases as the challenge deviates from the ideal.

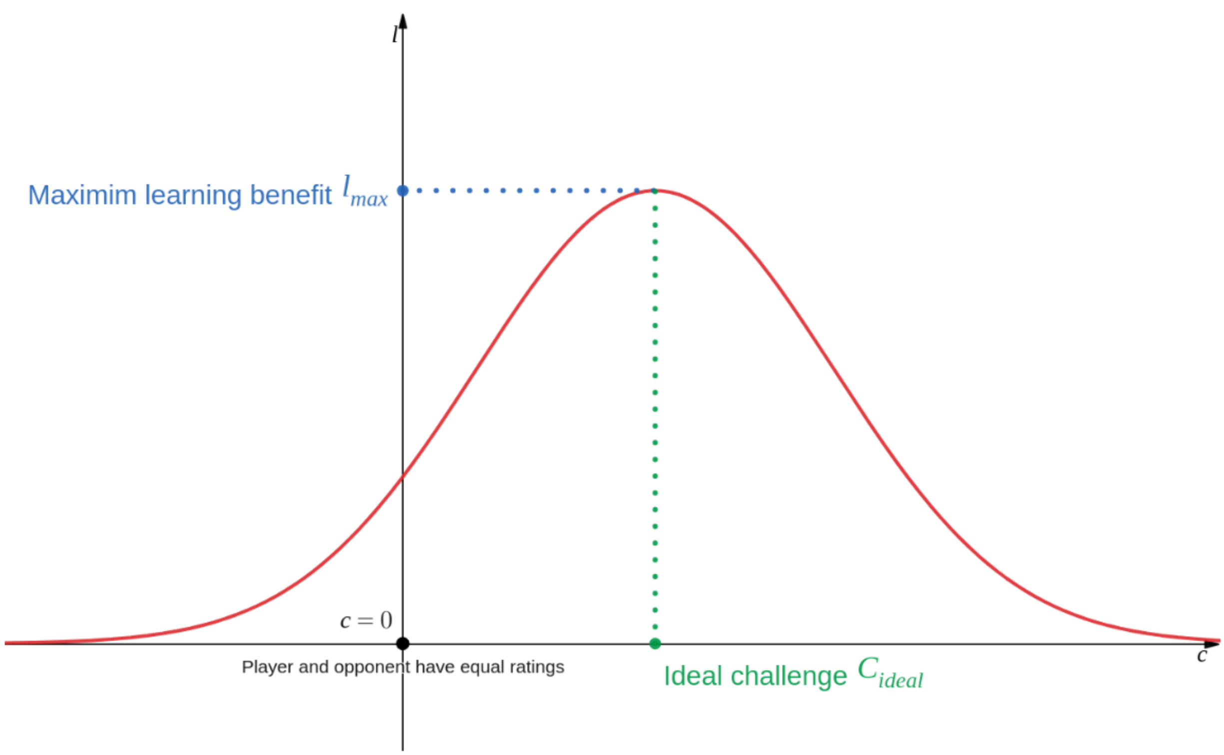

Figure 3.2: Learning benefit vs. rating difference with opponent

In Eq. (3.1), the learning reward L from previous games can then be defined as the mean of the learning benefit l over those games. That is, if we take the last N games into account, then the intrinsic rating I of a player at time t is defined as

$$I_t = R_t + < l >_{Nt} = R_t + \frac{\sum_{k=1}^{N} l_{max} e^{-(c_{t-k} - C_i)^2}}{N} \tag{3.5}$$

where $< l >_{Nt}$ is the mean learning benefit of the last N games, previous to time t.

Taking only the learning history from the last N games into account gives the model the benefit of letting I ultimately become R if the player is not facing challenges in recent games. This makes sense because old learning benefits from past challenges should already manifest in current match results and be reflected in the official rating.

Intrinsic rating, reflecting challenge and the learning that can occur by it, affects performance and ultimately official rating following the process in Figure 3.3.

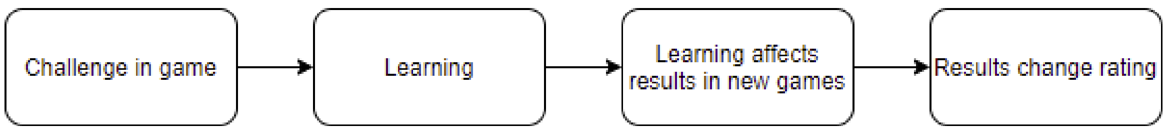

Figure 3.3: Flow of how challenge ultimately changes rating

3.7.2 Game outcome simulation

For a player A, the probability P_A of winning a game against player B can be calculated through the same method used by FIDE to calculate the expected outcome of a game, but using intrinsic ratings instead of the official ones:

$$P_A = \frac{1}{1 + 10^{(I_B - I_A)/400}} \tag{3.6}$$

This is intended to account for the impact of learning acquired in previous games in the outcome of the game. The effect can be seen more clearly if one considers two players of the same rating playing against each other. In the usual ELO model, they both have the same expected outcome of 0.5. In the model proposed, the player that has acquired the highest learning by facing higher challenges in the past games will have a higher probability of winning.

Every game has a winner, there are no draws in the ABM. This decision is made to build a simpler model that still allows us to explore emergent relations between segregation and improvement, and to avoid introducing additional arbitrary parameters to determine when a game outcome is a draw. This means that the probability of a win by player B is $P_B = 1 - P_A$.

Every time a game is played in the ABM, new ratings for the two players involved are calculated as well as their respective learning benefits acquired. It is important to note that learning benefits from a game, computed with Eq. (3.4), are independent of the game result and they can be calculated either before or after the game result is simulated.

After a game outcome is simulated, new ratings for both players are calculated by the usual ELO formula employed by FIDE

$$\Delta R_A = K_A \left(S_A - E_A\right) \tag{3.7}$$

where

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{3.8}$$

is the usual expected outcome using only ratings; no learnings or intrinsic ratings are involved in the calculation of new ratings.

Learning in this manner is a process that affects the outcome of games, but, as in reality, is itself invisible to the eyes of the system organizing tournaments and calculating ratings.

3.7.3 Tournament assignment

To assign players to tournaments we begin by creating two available tournaments: one open tournament and one women-only tournament. Players are assigned to either of these tournaments by order of rating, highest ratings first. If the player with the highest rating is in the *Men* group, he enters the available open tournament; if she is a player in the *Segregated* group, then she enters the available women-only tournament; if she is a player in the *Test* group, then she chooses the type of tournament. The choice of the player in the *Test* group is modeled as a random selection with probability P_{seg} of being assigned to the women-only tournament and probability $1 - P_{seg}$ of being assigned to the open tournament.

After this, the next player with the highest rating is assigned to a tournament following the same criteria. This is repeated until all players have been assigned to a tournament. When a tournament reaches a number of players equal to the Standard tournament size parameter (100 in our simulations) no more players can be assigned to it and a new tournament of the same type is created and made available. The only exception is the last tournament created - for the lowest rated players. This tournament may have between 50 and 149 players. This is done to avoid having a tournament with too few players.

3.7.4 Tournament play

Tournaments in the ABM are played in a *Swiss-system* format. In this format, all players in the tournament play the same fixed number of games (rounds). In each round, game pairings are arranged so that opponents have the most similar number of previous wins in the tournament. Players can only play once against each other in a tournament. In the first round, pairings are based on rating arranging similar players together. Chess tournaments are usually in the Swiss-system format to avoid eliminating players and because the *round-robin* tournament format in which each player plays every other player at least once is only practical for a small number of players. In the ABM, the number of rounds played in each tournament is given as a parameter.

Chapter 4

ABM Simulations

Several simulations of the model and algorithm were run using Netlogo, with various values of the population parameters and fixed values of tournament parameters as well as fixed values for the parameters of the learning model (Table 4.1).

Population composition parameters	
Number of players	2025
Women's participation rate P_w	0.08, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7
Segregation preference of the Test group P_{seg}	0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
Tournament parameters	
Standard tournament size	100
Number of rounds	10
Learning model parameters	
Ideal challenge	200
Maximum learning benefit	200
Benefit spread	200
Number of games in learning history	30

Table 4.1: Parameter values used in the ABM simulations

Twenty simulations were run independently for every combination of possible values of the P_w and P_{seg} parameters. In total, 12 320 simulations were run. Each run went until step 500, which will be referred to as *last step*.

At each step i, the following outcome values of the simulated system of players were recorded:

- MenTop20_i: Mean rating of the Men's top 20 players by rating
- TestTop20_i: Mean rating of the Test group's top 20 players by rating
- SegregatedTop20_i: Mean rating of the Segregated group's top 20 players by rating

From these values we calculated descriptors of the performance difference between groups:

- Test-Segregated$_i$ = TestTop20$_i$ − SegregatedTop20$_i$ − [TestTop20$_0$ − SegregatedTop20$_0$]. The initial rating difference between the groups was subtracted from each value. This is done to make this variable represent only the rating difference created by the system's evolution, not by the initial random sampling.

- Men-Test$_i$ = MenTop20$_i$ − TestTop20$_i$. Note that in this case we don't subtract the initial difference between these two groups. This is because this initial difference is actually a product of a difference that exists in reality, and not created by randomness in our model.

Variable names reported with an asterisk (Test-Segregated*, for example) will indicate that the mean of the 20 simulations with the same combination of parameters is being reported. This is possible for variables such as the Mean rating of the top 20 players because each mean is taken over the same number of values (20) and therefore each top 20 mean rating has the same weight when averaged with others.

Although we examined results for different values of all the parameters, some cases are of special interest: the cases with zero segregation preference –when the Test group's players always play in open tournaments– and the cases with 0.08 participation rate –the true participation rate of women at the beginning of 2010.

The choice of 200 as a value for the model parameters is somewhat arbitrary, but not gratuitous. Regarding Ideal Challenge and Benefit Spread: if in a game between two players of equal rating one of them has been playing only games against opponents 200 points better, and the other has been playing only games against opponents of equal rating, then, in terms of intrinsic rating given by Eq. (3.1), the player that has been facing more challenge will be expected to perform as if having about 126 rating points more than the player that has been facing opponents of equal rating. In such a game, using Eq. (3.6), instead of both players having expected outcomes of 0.5, the expected outcomes would be 0.67 and 0.33 in favor of the player that has been facing the greater challenge. This advantage obtained from that level of challenge is not unreasonable, but also not negligible. As for Benefit Spread, a value of 200 means that the curve in Figure 3.2 is sufficiently wide so that the difficulty doesn't have to be too close to the Ideal challenge to obtain non-negligible learning, but also not too wide as to give almost the same benefit for very different challenge values.

4.1 Results

The rating evolution of the Test, Segregated and Men groups can be visualized to evaluate the evolution of the performance gap between groups for different values of the population parameters. For instance, in Figure 4.1 (left) we follow the Mean ratings of the three groups for zero Segregation preference of the Test group and 0.08 Women's participation rate. We can see a clear differentiation between the Test and Segregated groups' top 20s performance: the test group, that plays open tournaments only, chases after the Men; while the Segregated group, that plays exclusively female tournaments, stagnates and lags behind. Also, the variance in the two groups of women is larger than the variance of rating of the men's top 20. An ANOVA test of the final rating of the Test and Segregated groups shows that their means are significantly different ($p - value < 2x10^{-16}$). We can compare this observation with the case where Segregation preference of the Test group = 1 (Figure 4.1, right). Now that the Test group has a Segregation preference of 1, just as the Segregated group, all women play together in all their tournaments, as in one unified group. As expected, we see that the Test and Segregated groups are practically indistinguishable throughout the simulation and their means are not significantly different (ANOVA $p - value = 0.965$).

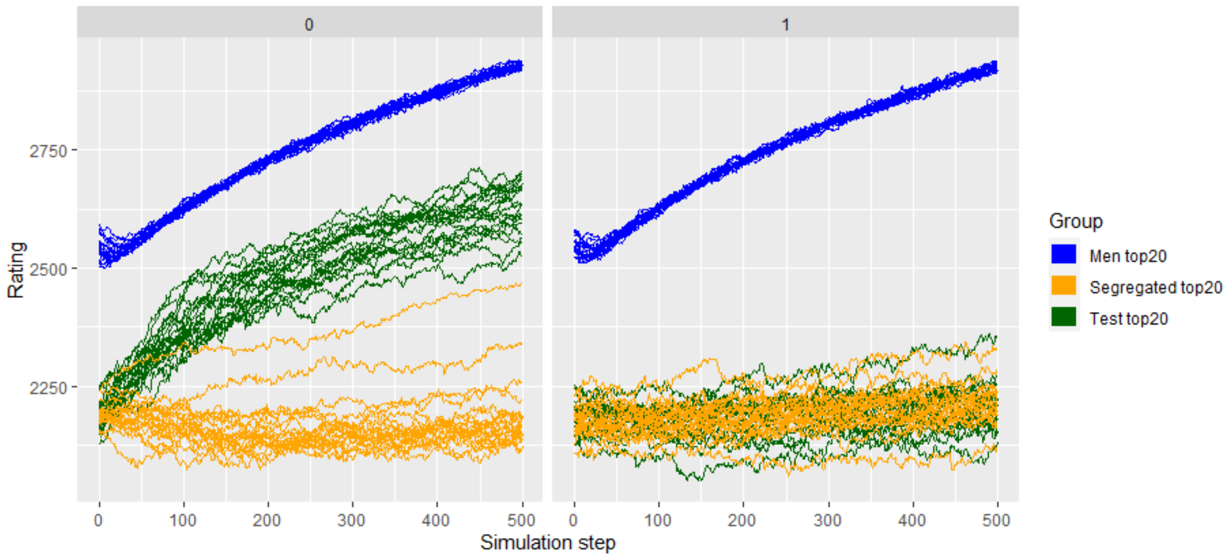

Figure 4.1: Rating evolution of the mean of the top 20 in each group when the Test group has zero segregation (left) and complete segregation (right). Women's participation rate of 8%, Ideal challenge of 200, Maximum benefit 200 and Benefit spread 200.

The same analysis can be done for all the different rates of women's participation (Figure 4.2). We see that at every participation rate, if they don't segregate at all, the Test group's elite performs better than the Segregated group. Also remarkable, for participation rates of 50% and higher the Test group's top 20 performs at a level similar to the Men's top 20. At a women's participation rate of 70%, the means of the Test and Men top 20s are not found to be significantly different (ANOVA p-value 0.791). Last, even the Segregated group performs better as women participation increases. These results show that the negative effect of self-segregation is bigger for minorities, and that even minorities that do not segregate themselves struggle to catch up.

We can also fix the participation rate and look at the effect that changing the segregation preference has on the groups' performance. In Figure 4.3 we plot the time evolution of the rating of each group's top 20 for a Women's participation rate of 8%. Each panel corresponds to a different Segregation preference of the Test group. The main observation here is how the advantage that the Test group can have over the Segregated one is almost lost when the Test group has a strategy of segregating in more than 40% of their games. In fact, ANOVA tests comparing the ratings of both female groups at the end of the simulation give no significant difference (with significance level of 95%) between the group's means for all simulated values of women's participation rates equal or greater than 60%. So mixing participation in open tournaments with female tournaments is not a viable strategy: female tournaments need to be actively avoided.

We can directly study the value of the difference between the top 20 of the two women's groups (Test and Segregated) for the extreme values of Segregation preference (0 and 1). In Figure 4.4 each line shows how the value of this difference changes with time in individual runs of the simulation. We are comparing the values of those differences on independent simulations, each of which has a Test group with a value of either 0 or 1 preference of segregation. We can see that in the case of total segregation (Segregation preference = 1) the difference between the two group's elite players stays consistently around 0 varying from -100 to 100. On the other hand, when the players in the Test group are not segregated at all, their advantage increases during the entire simulation (except for local variation), reaching values up to 5 times greater than in the case of complete segregation.

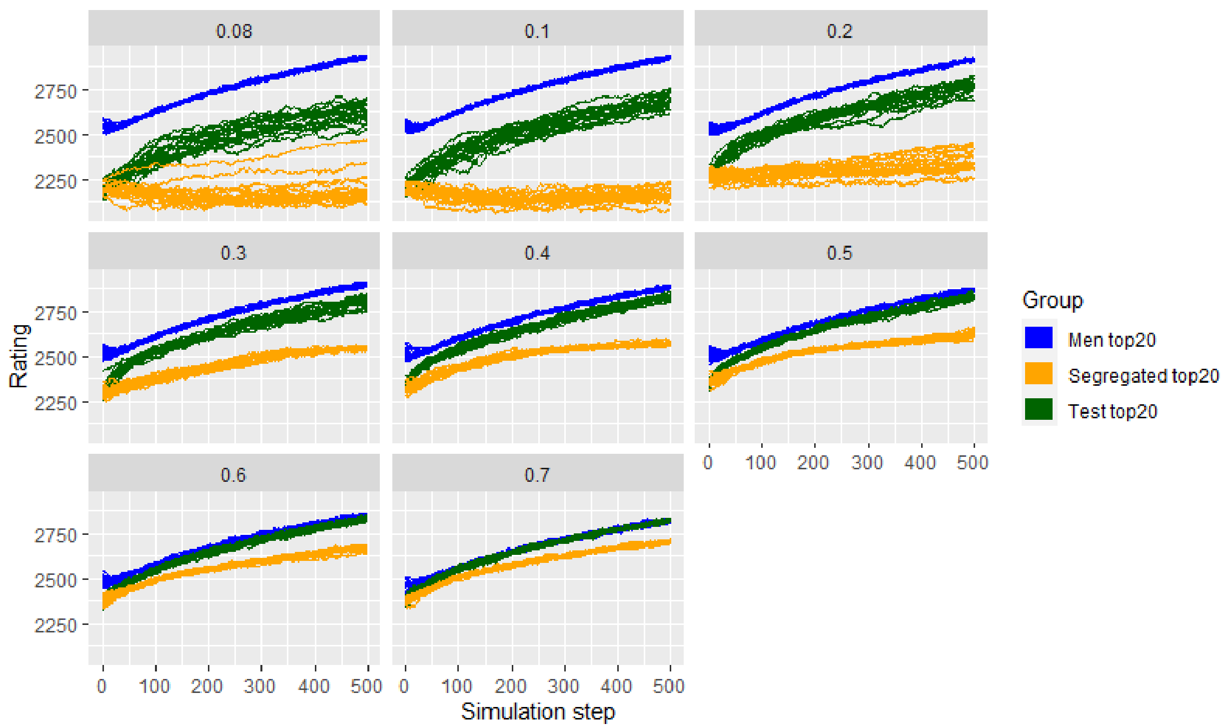

Figure 4.2: Rating evolution of each group's top 20 for different women's participation rates. Ideal challenge, Max benefit and Benefit spread are 200. Segregation preference of the Test group is 0.

We want to get a sense of how both the segregation preference of the Test group and the participation rate of women in the population relate to the groups' performance. Figure 4.5 shows Test-Segregated top 20's difference at the last step of the simulation for various values of the Segregation preference parameter. We can see that the difference between the groups' top 20s is almost exclusively in favor of the Test group (positive values on the y-axis). The advantage obtained by the Test group over the Segregated group is much larger when the Test group has the lowest values of Segregation preference. As we can see by the color code representing the Participation rate of women in the population, the greatest advantages in favor of the Test group at low Segregation correspond to the lowest women's participation rates (20% or less). When the Segregation preference of the Test group is below 0.5 (they play half their games in women-only tournaments), the advantage they gain over the Segregated group depends greatly on the fraction of the population that is female. On the contrary, if the Test group is segregated above 50%, they gain little or no advantage over the Segregated group and women's participation rate makes almost no difference. We thus confirm that self-segregation is only an issue for minorities, and that in such case even a small trend towards self-segregation harms performance.

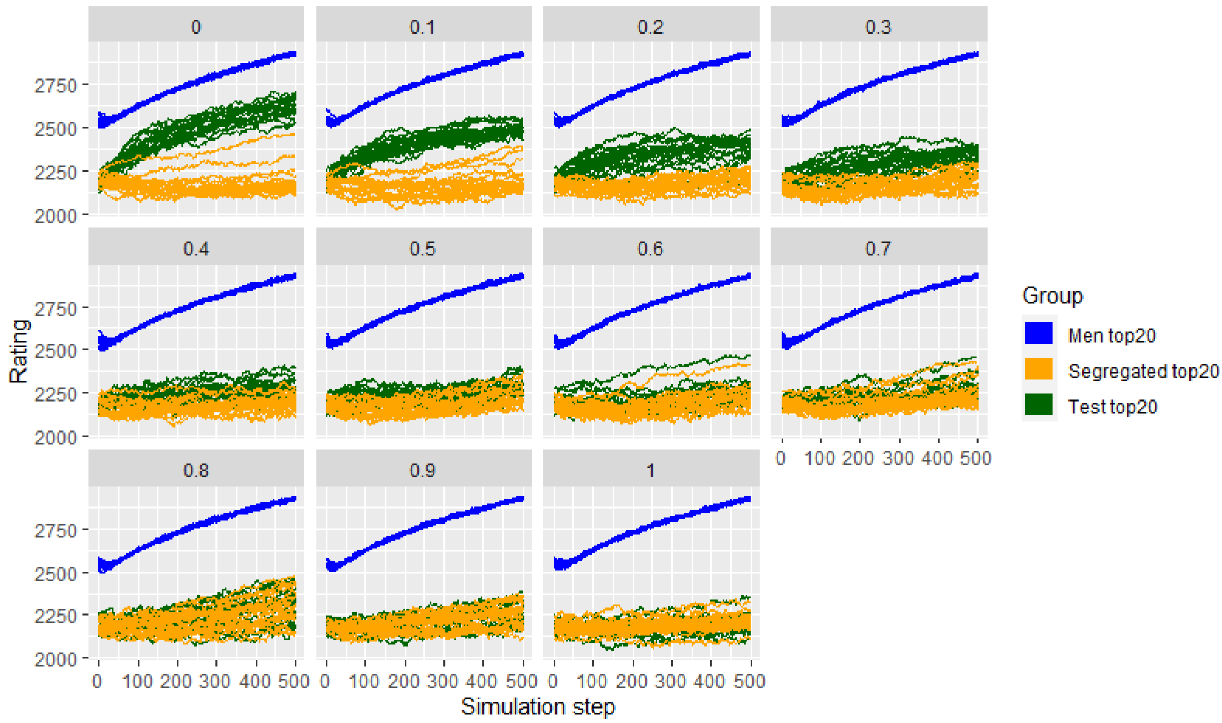

Figure 4.3: Rating evolution of each group's top 20 for different values of Segregation preference of the test group. Women's participation rate is 8%, Ideal challenge, Max. benefit and Benefit spread are 200.

Men's top 20 advantage over the Test's top 20 is also very dependent on women's participation rate (Figure 4.6). For most levels of women's participation rate, the best strategy for the Test group is to enter in competition with the dominant group (more open tournaments). However, if participation levels are very high, above 60%, it is actually better for the women in the Test group to segregate themselves completely (always play in women-only tournaments). This finding, though surprising at first, has a reasonable explanation: at these high levels of women's participation there are less men, so the initial sampling of men from the FIDE list will produce less male players near the top rating, resulting in lower initial top 20 ratings for the men. The opposite effect is happening at the same time in the bigger population of women, more players will be sampled near their highest rating, giving the Test group (and also the Segregated group) higher initial ratings of their top 20. This is confirmed by the average initial values of Men-Test: at 8% of women participation Men's top 20 starts with 369 points of advantage, while at 70% of women's participation rate Men's elite starts with only 83.2 rating points above the Test's top 20. Also, when having a larger population, the women will have a greater chance of producing more outliers in the course of the simulation.

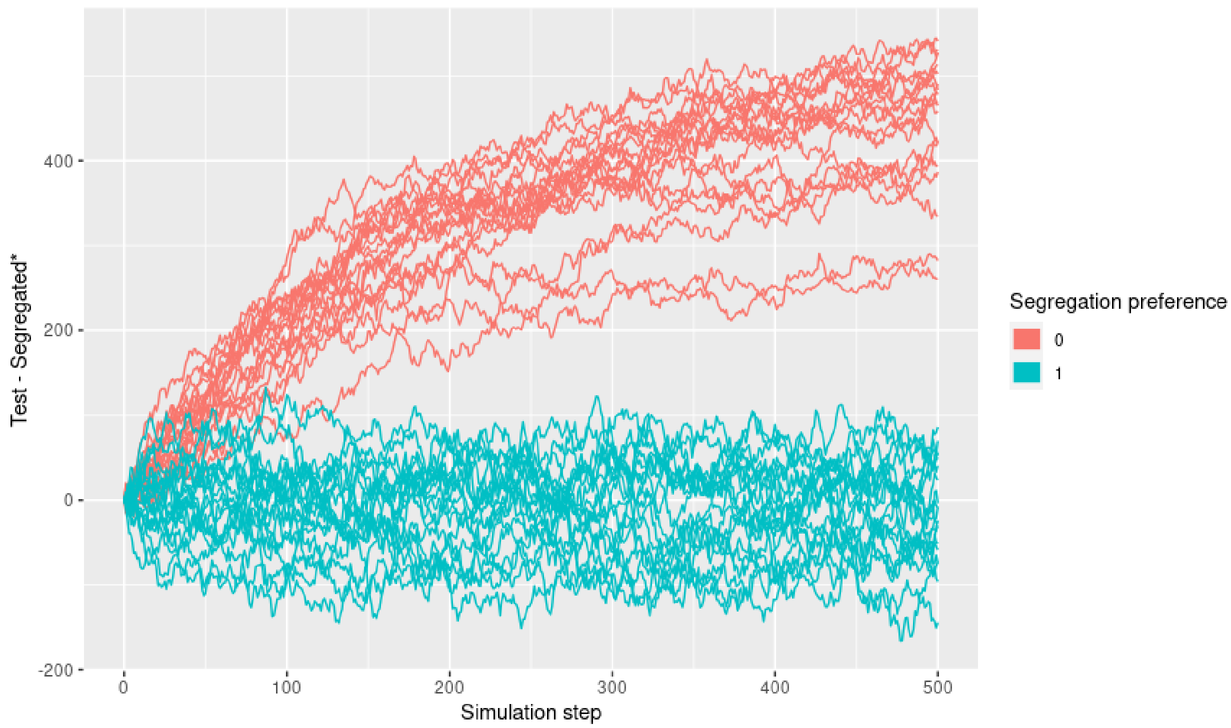

Figure 4.4: Evolution of the difference between Test and Segregated groups' top 20 for values 0 and 1 of Segregation preference of the Test group. Women's participation rate is 8%. Ideal challenge, Max benefit and Benefit spread are 200.

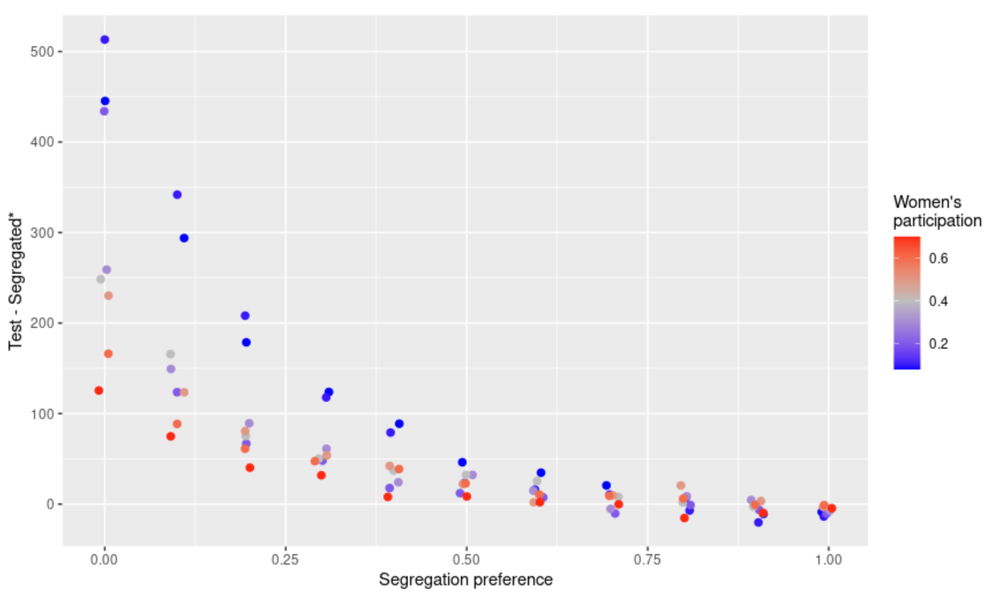

Figure 4.5: Final difference between the Test and Segregated group's top 20s vs. Segregation preference of the Test group for different participation rates of women. Ideal challenge, Max benefit and Benefit spread are 200.

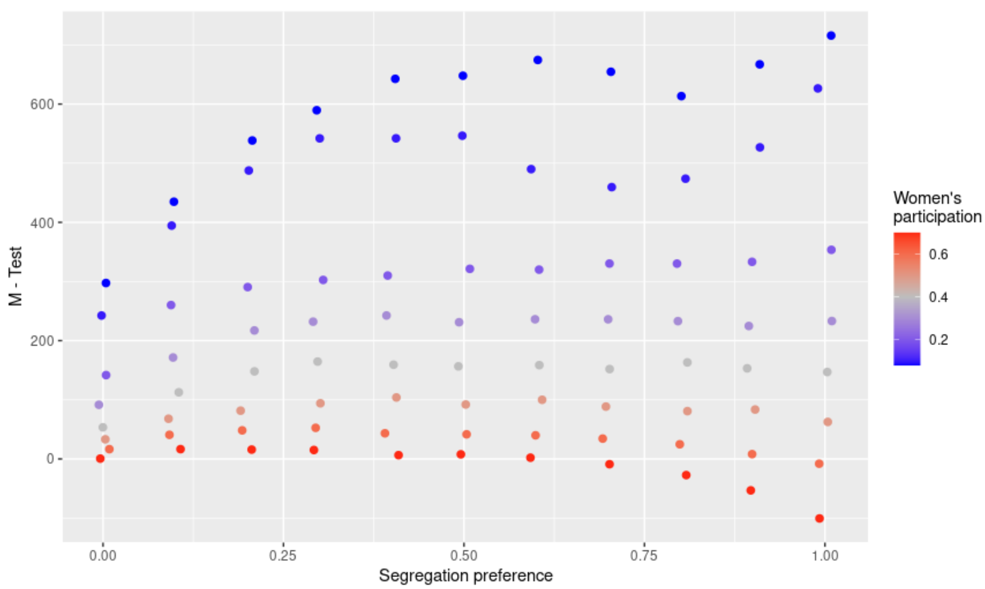

Figure 4.6: Final difference between the Men and Segregated group's top 20s vs. Segregation preference of the Test group for different participation rates of women. Ideal challenge, Max benefit and Benefit spread are 200.

Chapter 5

Analysis of the Parameter Space of the Learning Model

To study the possible effects that the parameters of the learning model can have on the results, we simulated the ABM with different values of Ideal Challenge, Maximum benefit and Benefit spread as follows:

- Ideal Challenge: 10, 50, 100, 150, 200, 250, 300, 350, 400
- Maximum benefit: 0, 50, 100, 150, 200, 250, 300, 350, 400
- Benefit spread: 10, 50, 100, 150, 200, 250, 300, 350, 400

Meanwhile, the population and tournament parameters, as well as the learning history parameter, were kept fixed with the following values:

- Number of players: 2025
- Women's participation rate: 0.08
- Segregation preference of Test group: 0
- Standard tournament size: 100
- Number of rounds: 10
- Number of games in learning history: 30

We ran 20 simulations of each combination of parameters for a total of 14 580 runs.

5.1 Results

In Figure 5.1 we show the final values of the Test - Segregated* top 20's difference for all simulated values of Maximum benefit, Ideal challenge and Benefit spread. Each point in the plot is the average value of Test - Segregated a the last step of the 20 runs that had the same values of the three learning parameters. We found that high Ideal challenge produces higher differences in favor of the Test group, specially when accompanied by high Benefit spread. Higher Maximum Benefit is also associated with a high difference in favor of the Test group but the effect is less significant after 300 points of Max. Benefit.

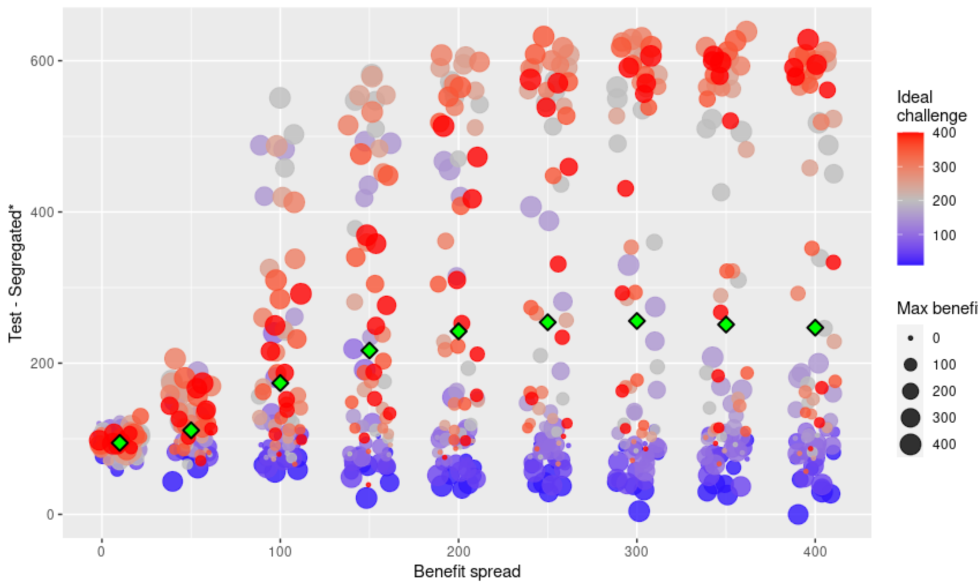

Figure 5.1: Test - Segregated at the final step vs. the Benefit Spread, Ideal Challenge and Max Benefit parameters. Women's participation rate is 8% and Segregation preference of the Test group is 0.

A different visualization of the exploration of the parameter space is shown in Figure 5.2. We observe that for low values of Max Benefit, as well as for low values of Benefit Spread, the values of the other parameters have no influence on the final difference between Test and Segregated groups, which will be minimal in every case. Also, for low values of Ideal challenge the values of the other two parameters have little importance and the outcome is always low Test-Segregated difference (leftmost points in every panel are always blue). This observation may be explained by the little difference that the segregation preference implies in this case, because if the rating difference that gives ideal challenge is very small, women can easily find them within their own group.

For parameter combinations of values from 100 to 300 in Benefit Spread and from 200 to 400 in Max Benefit, an initial increase in Ideal Challenge corresponds to increase in the Test-Segregated difference, but further increase then corresponds to decrease in the difference (though always in favor of not segregating). That is, for models with values of Max Benefit and Benefit spread in the mentioned ranges, there are corresponding optimal values of Ideal Challenge. This is consistent with what we saw in the simulations ran with fixed Maximum benefit (200) and Benefit spread (200). One possible explanation is that when the Ideal challenge is set too high players can't get the highest learning benefits because they can't find the opponents that give them so much challenge: opponents would have to be so many points above them that they either don't exist or they can't be found in the player's tournaments. In general, the highest Test-Segregated differences are found for values of Ideal Challenge in the range from 200 to 350 rating points.

Also notable is that, for the combinations Max Benefit > 200 and Benefit Spread > 100, there are mainly two distinct groups of possible Test-Segregated outcomes: low and high. This suggests that for those combinations of those two learning parameters the system has a threshold behaviour near certain values of Ideal challenge around which the outcome changes very rapidly.

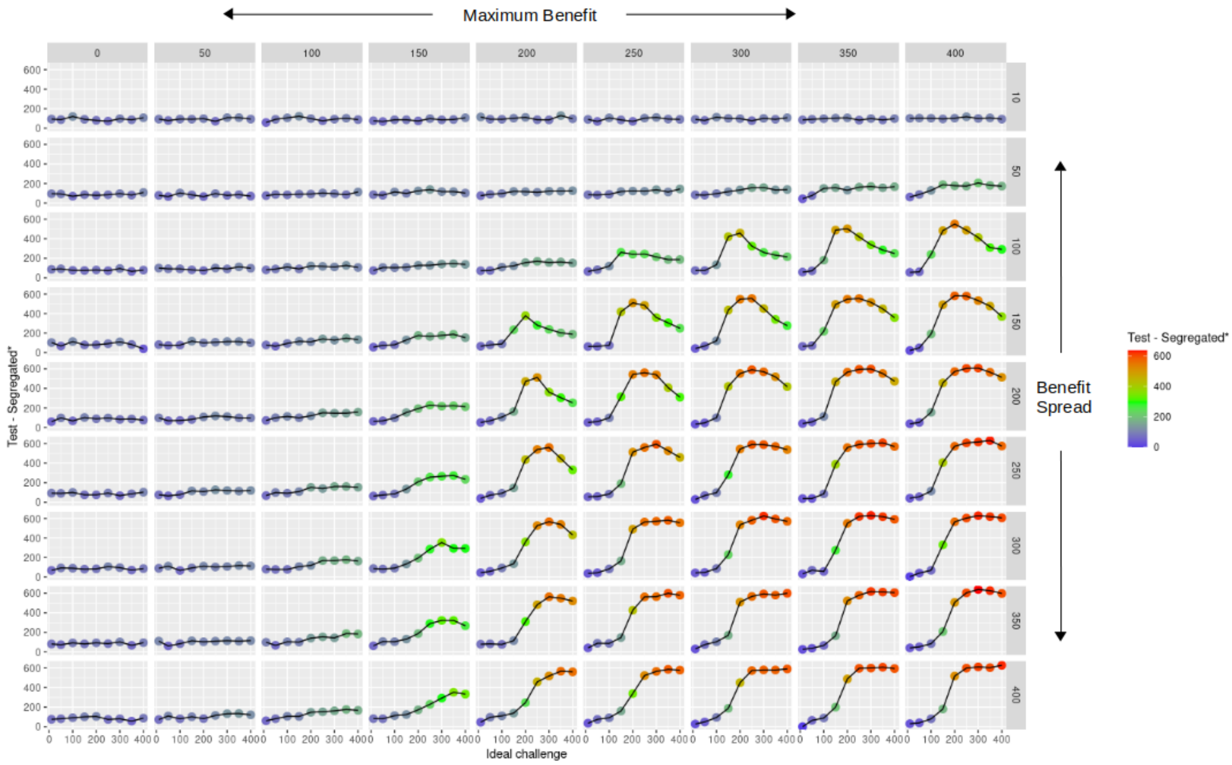

Figure 5.2: Test - Segregated at the final step vs. Ideal challenge for all the combinations of Max Benefit (columns) and Benefit spread (rows).Women's participation is 8% and Segregation preference of the test group is 0. Test - Segregated is also color-coded to make comparison between rows easier.

Experiments with a Maximum Benefit parameter of 0 are of particular interest. The value of zero for this parameter is equivalent to running the simulation with the learning model turned off. This cases provide a kind of control experiments to which the other results can be compared. One key observation from these experiments with zero Maximum Benefit, visible in Figure 5.2, is that the final values of Test - Segregated* are not centered around zero in these cases. This means that even without learning, the Test group is getting a performance advantage associated with not segregating. Since the payers in the Test group, specially the elite, have the chance to face the best male players in the population, on a few of those times they may get to win against the odds, gaining big rating points from those encounters that the Segregated group never gets the opportunity to have.

Another analysis that the experiments with zero Maximum Benefit allow us to do is to judge the relative effect that learning has on the observed group differences when it is non-zero. By comparison, the values of Test - Segregated* at the end of simulation are around six hundred rating points in the highest cases, about six times more that the control value with Maximum Benefit of zero. From this we can infer than the observed group differences analyzed in Chapter 4 are indeed a consequence of the learning acquired by players in the ABM.

Finally, a notable finding is the fact that all combinations of parameters show a final difference in favor of the Test group over the Segregated one. This is more relevant in the experiments in which the learning model shows effects different from the control experiments. This is important because it shows how segregation can produce a negative impact in group performance due to learning mechanisms, while taking some of the weight off the particular parametrization of the learning model used to show the effect.

Conclusions

In this work we studied how participating in games against challenging opponents may help players improve their performance. In the specific dominion of chess, in which women have very low representation at the highest ratings, we wondered if self-segregation of women into women-only tournaments could cause them to miss opportunities to improve their performance through learning from facing more challenging male players.

Data from FIDE-rated games during the last decade reveal that high-rated female players play predominantly in women-only tournaments. Also, a correlation was found between level of challenge and rating improvement for both sexes: high challenge correlates positively with rating improvement, while low challenge correlates negatively. In the period of study, the group that improved the most was also the group that challenged themselves the most. This gives credit to a central idea of the theory of desirable difficulties in which future performance is improved by facing difficult tasks.

Through an implementation of the theory of desirable difficulties in an agent-based model of competitive chess, we have shown that women-only tournaments can have a negative impact in the performance gap between women and men at the highest level of play. Indeed, performance improvement correlated negatively with the fraction of women-only tournaments played. This happened because self-segregation from the higher-rated male players deprived women from opportunities to face challenges that improve their mid-term learning of the game and therefore their performance in the chess ratings. Female players in the model who opted to play open-tournaments only, on the contrary, learnt from these challenges and visibly closed the gap separating them from men.

This effect was found to be stronger when minorities are involved. At the actual participation rate of 8%, women who segregated themselves into women-only tournaments stagnated in the ratings, while men and women who played open tournaments constantly improved. As women's participation was increased in the simulations, self-segregated women improved their ratings further and further, though they still performed below men and women playing open tournaments. Above 50% participation rate, women playing open tournaments caught up to the level of male players.

Notably, the qualitative aspect of the simulation results, i.e., that segregation can produce a comparatively negative effect on the improvement of the elite of the segregated group, were found to hold under a wide range of parameters of the learning model. This is one argument in favor of suggesting that the specific details of the learning process are not as important to reach a conclusion in favor of segregating less, as the assumption that learning in competition and it's model fulfill general requirements of desirable difficulties.

These results suggest that closing the rating gap between men and women in the highest-rated levels in chess requires closing the participation gap as well as watching segregation levels to pay the smallest possible learning cost. Eliminating self-segregation completely would be the first impulse. However, self-segregation of women in chess exists for many reasons, some of them including the promotion of women's chess. For instance, it can be intimidating for girls to be such a small minority in open children and youth tournaments; this is an additional emotional hurdle for them that the boys have the advantage of not having to overcome. Also, the existence of women-only tournaments means there are women champions that can be role models, attracting young girls to the game. It can then be argued that women-only tournaments are an important measure towards closing the participation gap. Remember that our model predicts that no learning cost would come from self-segregation if women weren't a minority. In a field of such low representation of women as today's chess, having some amount of women-only tournaments is probably a good idea.

With these considerations we can say, with the idea of improving both participation and performance, and much in the spirit of desirable difficulties, that it's fine to have some self-segregation in women chess, but probably not as much as we see today.

Bibliography

Abar, Sameera et al. (2017). “Agent Based Modelling and Simulation tools: A review of the state-of-art software”. In: *Computer Science Review* 24, pp. 13–33.

Assenza, Tiziana, Domenico Delli Gatti, and Jakob Grazzini (2015). “Emergent dynamics of a macroeconomic agent based model with capital and credit”. In: *Journal of Economic Dynamics and Control* 50, pp. 5–28.

Axelrod, Robert (2006). “Agent-based modeling as a bridge between disciplines”. In: *Handbook of computational economics* 2, pp. 1565–1584.

Bilalić, Merim, Peter McLeod, and Fernand Gobet (2007). “Personality profiles of young chess players”. In: *Personality and Individual Differences* 42.6, pp. 901–910.

Bilalić, Merim, Kieran Smallbone, et al. (2009). “Why are (the best) women so good at chess? Participation rates and gender differences in intellectual domains”. In: *Proceedings of the Royal Society B: Biological Sciences*. ISSN: 14712970. DOI: 10.1098/rspb.2008.1576.

Bjork, Elizabeth L, Robert Bjork, and Mark A Mcdaniel (2011). “Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning”. In: *Psychology and the real world: Essays illustrating fundamental contributions to society* 2, pp. 55–64.

Bjork, R. A. (1994). “Memory and metamemory considerations in the training of human beings”. In: *Metacognition: Knowing about knowing*.

Blanch, Angel, Anton Aluja, and Maria Pau Cornadó (2015). “Sex differences in chess performance: Analyzing participation rates, age, and practice in chess tournaments”. In: *Personality and Individual Differences*. ISSN: 01918869. DOI: 10.1016/j.paid.2015.06.004.

Bosse, Tibor et al. (2015). “Agent-based modeling of emotion contagion in groups”. In: *Cognitive Computation* 7.1, pp. 111–136.

Bruch, Elizabeth and Jon Atwell (May 2015). “Agent-Based Models in Empirical Social Research”. In: *Sociological Methods and Research* 44.2, pp. 186–221. ISSN: 15528294. DOI: 10.1177/0049124113506405.

Chabris, Christopher F. and Mark E. Glickman (2006). “Sex Differences in Intellectual Performance”. In: *Psychological Science*. ISSN: 0956-7976. DOI: 10.1111/j.1467-9280.2006.01828.x.

Charness, Neil and Yigal Gerchak (1996). “Participation rates and maximal performance: A Log-Linear Explanation for Group Differences, Such as Russian and Male Dominance in Chess”. In: *Psychological Science* 7.1, pp. 46–51. ISSN: 09567976. DOI: 10.1111/j.1467-9280.1996.tb00665.x.

Charness, Neil, Michael Tuffiash, et al. (2005). “The role of deliberate practice in chess expertise”. In: *Applied Cognitive Psychology*. ISSN: 08884080. DOI: 10.1002/acp.1106.

Cioffi-Revilla, Claudio (2014). "Introduction to computational social science". In: *London and Heidelberg: Springer*.

Clark, Courtney M and Robert A Bjork (2014). "When and why introducing difficulties and errors can enhance instruction". In: *Acknowledgments and Dedication*, p. 20.

Conte, Rosaria et al. (2012). "Manifesto of computational social science". In: *The European Physical Journal Special Topics* 214.1, pp. 325–346.

Dyble, Mark et al. (2015). "Sex equality can explain the unique social structure of hunter-gatherer bands". In: *Science* 348.6236, pp. 796–798.

Epstein, Joshua M (2008). "Why model?" In: *Journal of Artificial Societies and Social Simulation* 11.4, p. 12.

Fair, Ray C. (2007). "Estimated age effects in athletic events and chess". In: *Experimental Aging Research*. ISSN: 0361073X. DOI: 10.1080/03610730601006305.

FIDE (2021a). *FIDE Rating Regulations effective from 1 July 2017*. URL: https://handbook.fide.com/chapter/B022017 (visited on 06/02/2021).

— (2021b). *FIDE Title Regulations effective from 1 July 2017*. URL: https://handbook.fide.com/chapter/B01Regulations2017 (visited on 06/02/2021).

Gobet, F. and G. Campitelli (2007). "The role of domain-specific practice, handedness, and starting age in chess". In: *Developmental Psychology*. ISSN: 0012-1649.

Grimm, Volker, Uta Berger, Finn Bastiansen, et al. (2006). "A standard protocol for describing individual-based and agent-based models". In: *Ecological modelling* 198.1-2, pp. 115–126.

Grimm, Volker, Uta Berger, Donald L DeAngelis, et al. (2010). "The ODD protocol: a review and first update". In: *Ecological modelling* 221.23, pp. 2760–2768.

Grow, André and Jan Van Bavel (2015). "Assortative mating and the reversal of gender inequality in education in Europe: An agent-based model". In: *PloS one* 10.6, e0127806.

Howard, Robert W. (2005). *Are gender differences in high achievement disappearing? A test in one intellectual domain*. DOI: 10.1017/S0021932004006868.

— (2014). "Gender differences in intellectual performance persist at the limits of individual capabilities". In: *Journal of Biosocial Science*. ISSN: 14697599. DOI: 10.1017/S0021932013000205.

Kapur, Manu (2016). "Examining Productive Failure, Productive Success, Unproductive Failure, and Unproductive Success in Learning". In: *Educational Psychologist*. ISSN: 00461520. DOI: 10.1080/00461520.2016.1155457.

Klimek, Peter et al. (2015). "To bail-out or to bail-in? Answers from an agent-based model". In: *Journal of Economic Dynamics and Control* 50, pp. 144–154.

Knapp, Michael (2010). "Are participation rates sufficient to explain gender differences in chess performance?" In: *Proceedings of the Royal Society B: Biological Sciences*. ISSN: 0962-8452. DOI: 10.1098/rspb.2009.2257.

Lomas, Derek et al. (2013). "Optimizing challenge in an educational game using large-scale design experiments". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 89–98.

Maass, Anne, Claudio D'ettole, and Mara Cadinu (2008). "Checkmate? The role of gender stereotypes in the ultimate intellectual sport". In: *European Journal of Social Psychology*. ISSN: 00462772. DOI: 10.1002/ejsp.440.

Macal, Charles M (2016). "Everything you need to know about agent-based modelling and simulation". In: *Journal of Simulation* 10.2, pp. 144–156.

Macy, Michael W. and Robert Willer (2002). "From Factors to Actors: Computational Sociology and Agent-Based Modeling". In: *Annual Review of Sociology* 28, pp. 143–166. ISSN: 03600572, 15452115. URL: http://www.jstor.org/stable/3069238.

Metcalfe, Janet (2011). "Desirable difficulties and studying in the Region of Proximal Learning". In: *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork*, pp. 259–276.

Phillips, Sean (2010). *Story @ Www.Npr.Org*. URL: https://www.npr.org/templates/story/story.php?storyId=129214019.

Polgar, Susan (2019). *Why is there a need for some all-girl's or women's chess tournaments?* URL: https://gamesmaven.io/chessdailynews/womens/why-is-there-a-need-for-some-all-girls-or-women-s-chess-tournaments-jchwvXHHk0SuodXKDF7FFQ.

Rai, Varun and Scott A Robinson (2015). "Agent-based modeling of energy technology adoption: Empirical integration of social, behavioral, economic, and environmental factors". In: *Environmental Modelling & Software* 70, pp. 163–177.

Railsback, Steven F. and Volker Grimm (2011). *Agent-Based and Individual-Based Modeling: A Practical Introduction*. ISBN: 9788578110796. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Roring, Roy W. and Neil Charness (2007). "A Multilevel Model Analysis of Expertise in Chess Across the Life Span". In: *Psychology and Aging*. ISSN: 08827974. DOI: 10.1037/0882-7974.22.2.291.

Shahade, Jennifer (2005). *Chess Bitch. Women in the Ultimate Intellectual Sport*. Siles Press.

Soderstrom, Nicholas C. and Robert A. Bjork (2015). "Learning Versus Performance". In: *Perspectives on Psychological Science*. ISSN: 1745-6916. DOI: 10.1177/1745691615569000.

Vaci, Nemanja and Merim Bilalić (2017). "Chess databases as a research vehicle in psychology: Modeling large data". In: *Behavior Research Methods*. ISSN: 15543528. DOI: 10.3758/s13428-016-0782-5.

Wang, Haizhong et al. (2016). "An agent-based model of a multimodal near-field tsunami evacuation: Decision-making and life safety". In: *Transportation Research Part C: Emerging Technologies* 64, pp. 86–100.

Wang, Zhihui et al. (2015). "Simulating cancer growth with multiscale agent-based modeling". In: *Seminars in cancer biology*. Vol. 30. Elsevier, pp. 70–78.