MANCHESTER 1824

The University of Manchester

Economics
Discussion Paper Series
EDP-1605

Gender, Competition and Performance: Evidence from Real Tournaments

Peter Backus
Maria Cubel
Matej Guid
Santiago Sanchez-Pages
Enrique Lopez Manas

October 2016

Economics
School of Social Sciences The
University of Manchester
Manchester M13 9PL

Gender, Competition and Performance: Evidence from real tournaments

Peter Backus
University of Manchester & Institut d'Economia de Barcelona (IEB)

Maria Cubel
University of Barcelona & Institut d'Economia de Barcelona (IEB)

Matej Guid
University of Ljubljana

Santiago Sánchez-Pages
University of Barcelona

Enrique Lopez Manas
Google Developer Expert

Abstract

There is a growing literature looking at how men and women respond differently to competition. We contribute to this literature by studying gender differences in performance in a high-stakes and male dominated competitive environment, expert chess tournaments. Our findings show that women underperform compared to men of the same ability and that the gender composition of games drives this effect. Using within player variation in the conditionally random gender of their opponent, we find that women earn significantly worse outcomes against male opponents. We examine the mechanisms through which this effect operates by using a unique measure of within game quality of play. We find that the gender composition effect is driven by women playing worse against men, rather than by men playing better against women. The gender of the opponent does not affect a male player's quality of play. We also find that men persist longer against women before resigning. These results suggest that the gender composition of competitions affects the behavior of both men and women in ways that are detrimental to the performance of women. Lastly, we study the effect of competitive pressure and find that players' quality of play deteriorates when stakes increase, though we find no differential effect over the gender composition of games.

Keywords: Competition, Gender, Stereotype threat, Chess.
JEL Codes: D03, J16, J24, J70, L83, M50.

Acknowledgments. The authors are grateful for helpful comments from Wiji Arulampalam, Sendhil Mullainathan and Imran Rasul as well as audiences at the University of Manchester, University of Kent, University of Alicante, University of Barcelona, University of Pittsburgh and the Petralia Applied Economics workshop. Cubel and Sanchez-Pages acknowledge funding from the Spanish Ministry of the Economy and Competitiveness research grant ECO2015-66281-P.

'They're all weak, all women. They're stupid compared to men. They shouldn't play chess, you know. They're like beginners. They lose every single game against a man. There isn't a woman player in the world I can't give knight-odds to and still beat.' *Bobby Fischer, 1962, Harper's Magazine*

'Chess is a mixture of sport, psychological warfare, science, and art. When you look at all these components, man dominates. Every single component of chess belongs to the areas of male domination.' *Garry Kasparov, 2003, The Times of London*

'Girls just don't have the brains to play chess.' *Nigel Short, 2015, The Telegraph*

1 Introduction

Despite extensive research, debate and policy interventions, gender differences in labor market outcomes persist. The unconditional gender wage gap is about 18% in OECD countries. Only 1 in 7 board members in European and US companies are women. Three traditional explanations for this phenomenon are discrimination, differences in ability and differences in preferences for jobs (Polachek, 1981; Goldin and Rouse, 2000; Black and Strahan, 2001). More recently, a growing interest has developed around a fourth explanation: Gender differences in competitiveness (see Niederle and Vesterlund (2011) for an excellent survey). The existing evidence suggests that women perform worse than men of the same ability in competitive environments (Gneezy, Niederle, and Rustichini, 2003), which may lead women to 'shy away from competition' (Niederle and Vesterlund, 2007). Given that good management practice dictates that managers ought to create competitive environments to increase productivity (Bloom, Propper, Seiler and Van Reenen, 2015), gender differences in competitiveness might help to explain the persistent gender differences in wages and the under-representation of women in high-powered jobs.

The majority of studies on gender and competition confine themselves to experiments both in the lab and in the field. In general, results in this literature suggest that women are more reluctant to enter into a competition, even when they are no less able than men (Vandegrift and Brown 2005; Niederle and Vesterlund, 2007; Gupta, Poulsen, and Villeval, 2013), and that women may be less responsive to competition than men (e.g. Gneezy and Rustichini, 2004). Any such differences seem to be social rather than innate (Gneezy, Leonard and List, 2009). However, as Croson and Gneezy (2009) point out, many unanswered questions remain, including the effect of gender composition on gender differences on competitive performance.

In this paper, we contribute to the existing literature on gender and competition by studying

a natural setting. We employ data from tens of thousands of expert chess games played by highly skilled and dedicated players. Our first finding is that female players obtain worse outcomes than male players of the same ability. Second, we show that this performance gender-gap is due to the gender composition of the match: two players of identical ability perform similarly in single gender games, whereas female players obtain worse outcomes than male players of the same ability in mixed gender games. We then move to the study of the mechanisms behind this effect. We use the method developed by Guid and Bratko (2006, 2011) to compute the quality of play of each player in each game they play. This differs from measures of ability based on win-loss records, as it allows us to measure how well a player plays during a particular game. We find that the effect of the gender composition of the game on the underperformance of women is driven largely by female players making larger errors when playing against males. This is in contrast with male players, who play equally well regardless of the sex of the opponent. We also find that, on average, men persist longer before resigning when playing against a woman, decreasing the points that a female player can expect to earn against a male opponent. These results suggest that inter-gender competition changes the behaviour of both men and women in ways that are detrimental to the performance of women. Lastly, we study the existence of gender differences in response to increased competitive pressure. We find that the quality of play diminishes in games with higher stakes, though we do not find compelling evidence of heterogeneity in this effect over the gender composition of games.

Chess is an ideal testbed for the study of gender differences in competitiveness for a number of reasons. First, it is one of the few sports, if not the only one, in which male and female players engage in head-to-head competition against one another. Second, as chess is ultimately a computational problem, performance is almost exclusively a function of effort and ability. Unlike in other games, say, poker, luck plays virtually no role. Third, there exists a standard and well-established metric of players' ability, the Elo rating (discussed in detail below), allowing us to control for the relative abilities of the competitors. These ratings are publicly observable. Consequently, players have a very good sense of where they stand relative to their opponent at the start of the game. Thus we can largely rule out one of the main explanations for gender differences in competitiveness, namely, gender differences in overconfidence (Niederle and Vesterlund, 2011). Finally, our data has records of every move in each game, not just its outcome, so we can observe the choices made by players and the circumstances in which those choices are

made. As chess is a computational exercise (discussed below), we can objectively assess players' quality of play by comparing their chosen moves with the preferred move of a powerful chess engine.[1] For most competitive environments, sports or otherwise, such counterfactuals cannot be calculated.

Expert chess is relevant to the study of gender because it shares several important features with high-powered jobs and competitive professional settings. First, like board rooms, expert chess is a domain in where women are severely under-represented. Women constitute only about eleven percent of mixed-sex tournament players and two percent of Grandmasters. Second, as the opening quotes demonstrate, negative stereotyping against females is pervasive in high level chess as it is in professional settings (Auster and Prasad, 2016). Third, female players in chess underperform their male counterparts. The average female player has a rating that is 15 percent lower than the average male player (Blalic, Smallbone, McLeod and Gobet, 2009). There is currently (August 2016) only one woman, Hou Yifan, ranked among the top 100 players and there has never been a female world-champion. This performance gap echoes the gender wage gap that persists even after controlling for potential confounders (Blau and Kahn, 2016) because there is no compelling evidence showing that men are innately superior chess players, as we discuss below. Lastly, top chess players, be they male or female, are individuals with high levels of cognition, determination, tenacity and dedication. Expert female players, like women in highly competitive professional environments, have selected into a male-dominated and very demanding environment. Given this selection, one might think it unlikely that we observe any gender differences in performance. However, we do. The fact that we observe gender differences among this very select group of people suggests that vulnerability to such gender effects should be prevalent, and probably stronger, in wider domains.

In the next section we discuss the literature which has studied gender differences in competitiveness in real settings and in chess in particular. In Section 3 we present the data. In Section 4 we present our main result and determine that, after controlling for ability, age and other factors, players' performance is affected by the gender of the opponent. In Section 5, we discuss the mechanisms behind this result. Section 6 concludes.

[1] As a matter of fact, during competitions, commentators already employ computers to learn which next move is a player's best and can recognize mistakes almost immediately. See "How computers changed chess", *The Conversation*, May 2013. Available at https://theconversation.com/how-computers-changed-chess-20772

2 Related Literature

Competition in the field

The majority of studies on gender differences in competitiveness confine themselves to laboratory and field experiments. Gneezy et al. (2003) conducted a laboratory experiment where subjects had to solve mazes on the Internet. They found that women performed worse than men when the payment scheme was competitive but not under piece-rate compensation. Gneezy and Rustichini (2004) confirmed this finding for Israeli children in running competitions. In these relatively artificial setups, the perceived gender-bias of the task plays an important role. Günther, Ekinci, Schwieren, and Strobel (2010) find that females perform better than males when the task is perceived as female-biased. Along similar lines, Shurchkov (2012) finds that females overtake men when performing a verbal task under low-time pressure. Iriberri and Rey-Biel (2016) show that omitting information about the gender of the opponent helps to mitigate the underperformance of women in competition, indicating that part of the problem is who women compete against, not simply that they compete at all. Although clearly valuable for their ability to control confounding factors, these experiments are far removed from real interactions between men and women and from real competitive stakes. The question that remains is whether differences in competitiveness observed in experiments persist in real settings.

Only a handful of contributions have addressed gender differences in competitiveness in natural settings, in part due to the difficulty of finding appropriate conditions (e.g. observable ability, men and women competing on equal footing). One exception is educational competitions, and admissions to selective programs in particular, where men and women compete in an equal footing. This permits the study of differential responses to varying degrees of competition. Örs, Palomino and Peyrache (2013) compare the results of the same group of male and female students in a less competitive high school national exam and in a very competitive exam for entry into a selective French business school. The performance of female students dominates the one of male students in the less competitive exam whereas the opposite holds in the more competitive exam. A similar picture emerges from the study by Morin (2015), who takes advantage of an educational reform in Ottawa which shortened high school by one year. This meant that two cohorts of students graduated in the same year thus increasing competition for university places. Morin finds that the average grades and graduation rates of male students increased relative to

those of females. But gender effects do not always appear in competitive environments. For example, Lavy (2013) studies the behavior of Israeli teachers who participated in rank-order tournaments that rewarded them with cash bonuses based on the tests scores of their classes. Teachers were competing with others within their field and school. Male and female teachers did not respond differently to this new payment scheme.

While informative about how men and women respond to competition, these studies do not, however, tell us whether it is competing that hurts women's performance or competing against men. This question is addressed in Antonovics, Arcidiacono and Walsh (2009), who use data from the TV trivia show *The Weakest Link*. In the final round of the show, two players compete against each other. The authors find that male contestants are more likely to answer a question correctly when they face a female contestant than when they face another male. The gender of the opponent does not influence the performance of female contestants. This is in contrast with our finding of female players making larger errors when facing a male opponent.

Chess

Chess has been studied by psychologists for years because it involves high order cognition (see Charness (1992) for an early survey). Chess has also become a recent object of interest for economists. The cognitive power of expert chess players, combined with the computational nature of the game, makes chess a natural candidate for the study of strategic sophistication (Palacios-Huerta and Volij, 2009; Levitt, List and Saddoff, 2011). More closely related to our analysis, Gerdes and Gränsmark (2010) explore gender differences in risk behaviour by using chess data. They measure risk according to whether the opening moves of a game are deemed 'aggressive' or 'solid' by a number of expert chess players. They associate 'aggressive' openings with risk taking. They find that females are on average two percent less likely to use an 'aggressive' opening than male players. Males are more likely to use 'aggressive' openings when playing against females. Females also have a tendency to use more 'aggressive' openings against females but only against female players with higher rating than themselves. However, the authors also find that 'aggressiveness' reduces the probability of winning regardless of the gender of the opponent. This finding ultimately falls short to explain the gender performance gap in chess.

Gränsmark (2012) explores gender differences in time preferences. Again using survey data

on expert chess players, he finds that males play shorter games on average and that they are willing to pay a higher price to end the game sooner by arranging a draw. Hence, Gränsmark (2012) concludes that males are more impatient than women. We return to the issue of game length in our analysis below.

Two papers have specifically explored the response of female chess players to the gender of the opponent. Using online games, Maas, D'ettole and Cadinu (2008) find no gender differences in outcomes when the sex of the opponent remains unknown. Compared to that benchmark, women perform more poorly when they know they are playing against a male opponent. When they falsely believe to be playing against a woman, gender differences disappear again. They use data from "rapid" chess games (15 minutes long) and they do not control for the Elo rating of the participants. Rothgerber and Wolsiefer (2014) use field data and find that females underperform when playing against a male opponent. They again use short games (30 minutes) played by elementary, middle and high school students. Although they have information on students' pre- and post-game ratings, their ability measure is not as reliable as the Elo rating.

The closest paper to ours is the recent working paper from de Sousa and Hollard (2015) who also look at the effect of inter-gender competition using data from chess tournaments. It is reassuring that they find a gender effect as we do; women perform worse when playing against male opponents. They go on to consider whether the effect diminishes with experience (only very slightly) and with the Gender Gap Index of the player's home country (not at all). Though their data set is considerably larger than ours in terms of the number of games, they do not study the mechanisms underlying the observed gender effect nor do they employ a within game quality of play measure as we do. They are therefore relatively limited in how far they can study the mechanisms underlying the estimated effect, the central focus of our paper. Moreover, they do not address the issue of the (conditionally) random assignment of the opponent's gender whereas we confront the issue of identification directly. We are also the first ones to characterize the bias caused by the measurement error in Elo ratings arising from the observed gender effect. In addition, we study gender differences in the effect on performance of competitive pressure as measured by the stakes of games.

3 Data

Chess players are rigorous data collectors who systematically codify information on games played in tournaments and share it in publicly available archives. They collect a great deal of game information: date of the game, event at which it was played, all moves made, the color (white or black) each player plays with, the players involved, the outcomes, the FIDE[2] registration number of each player, which allows us to link the game data to information on their gender, age and affiliated national federation, and the ability of each player at the start of each game as measured by the Elo rating. Game data are generally stored in Portable Game Notation (PGN) files which can be read by chess programs allowing players to review how a particular game unfolded.

We take our data from the weekly publication "The Week in Chess" (TWIC). Every Monday, TWIC publishes game data from the largest and most notable tournaments from around the world. We use the PGN files published by TWIC for 2012 and the first six months of 2013 giving us information from 79,242 games played by 14,056 players from 154 national federations.[3]

Our data set is constructed by randomly selecting a player from each game (*white* or *black*). The selected player is our unit of observation, i or the 'player', and we denote the other person as the 'opponent'. Arranging the data in this way means we construct a panel of player i over games g. We can thus control for player i fixed effects to consider the effect of within i variation in game conditions, including the gender of i's opponent, on i's performance.

We restrict our sample in a number of ways. Following Gränsmark (2012) and Gerdes and Gränsmark (2013), we focus on expert chess players and drop those games in which i has an Elo rating less than 2000 (we keep games in which the opponent has an Elo less than 2000, though all our results are robust to their exclusion). Players at this level are regarded as experts and have generally committed between 2500 and 7500 hours of alone study (excluding coaching and group study) to achieve this level of skill (Hambrick, Oswald, Altmann, Meinz, Gobet and Campitelli, 2014). We also drop games which lasted fewer than 15 moves as we need games at least that long to compute our quality of play variable (details below). We also exclude players who play only one game in our sample. Our full analytic sample is therefore comprised

[2] FIDE stands for the Fédération Internationale des Échecs or World Chess Federation which is the international governing body of chess.

[3] These were the TWIC data available when we started working on this paper.

of 57,936 games played by 7,932 players. We define a second sample by excluding those players who only play against one gender in our sample (all male or all female opponents) reducing our sample to 28,759 played by 2,506 players, a sub-sample of players we call 'switchers'. We present descriptive statistics for male and female players in both the full and restricted sample of 'switchers' in Table 1.

Columns (1) and (2) are for the full sample of male and female players respectively, and columns (3) and (4) are for the sub-sample of switchers**.** In our full sample, male players earn an average of 0.53 points per game (the standard point system assigns one point for a win, 0.5 for a draw and zero for a loss) and female players earn 0.50, a differential which holds for switchers.[4] Players and opponents are about 31 years old on average, though females tend to be a bit younger. The degree to which chess is male dominated is apparent with 87 percent of the players in our full sample and 80 percent in our restricted sample being male. Note also that male players are more likely than female players to face a male opponent, an important point we return to below.

The Elo rating

The availability of the Elo ratings, created by the physicist Arpad Elo (Elo, 1978), is one of the major advantages of using chess data. The Elo is a cardinal rating of each player's ability as a function of the outcome of previous games played and the difference between the player's own Elo rating and that of the opponent in those games. The Elo rating has a minimum score of zero and no upper bound. For each 200 point interval from 100 to 1999 players are rated J to A. Players with Elo ratings of 2000-2199 are classified as 'Experts", 'Candidate Masters' have ratings in excess of 2200, 'International Masters' ratings above 2400 and 'Grand Masters' have ratings larger than 2500. There are also women's equivalent titles, though the Elo rating thresholds are 200 points lower than for the men's titles and some top female players have opted not to take on such gendered titles. Very few top players have obtained ratings of over 2700. The current world champion, Magnus Carlsen, achieved the highest Elo score ever obtained by

[4] Note that the expected points for players in our sample exceeds 0.5 as we condition our sample to include players with Elo ratings of at least 2000, while opponents may have an Elo rating below 2000.

a human player, 2882, in 2014. Top computer chess engines have Elo scores above 3200.[5]

A player's Elo rating at the end of game g can be expressed as

$$Elo_{ig+1} = Elo_{ig} + K\left[p_{ig} - E[p_{ig}]\right], \tag{1}$$

where ELO_{ig} is i's Elo rating at the start of game g, p_{ig} is the points the player obtains in game g, and K is an adjustment parameter. Following the FIDE rules applying to our sample, K =15 if the player has a rating lower than 2400 and $K = 10$ once a player achieves a rating of 2400, even if her rating falls back below that threshold.[6] Finally,

$$E[p_{ig}] = \frac{1}{1 + 10^{\left(\frac{Elo^{o}_{ig} - Elo_{ig}}{400}\right)}}, \tag{2}$$

is the so called 'Elo curve', that is, the points i is expected to earn from a game against an opponent with an Elo rating of Elo^{o}_{ig} at the start of game g. As it can be seen, Elo points are earned (lost) by performing better (worse) in a game than the Elo curve predicts. Players always increase their Elo rating following a win and decrease it following a loss. The effect of a draw is positive if $Elo_{ig} < Elo^{o}_{ig}$, negative if $Elo_{ig} > Elo^{o}_{ig}$ and neutral if $Elo_{ig} = Elo^{o}_{ig}$.

The advantage of the Elo rating is two-fold. First, regardless of the player's subjective assessment of her ability relative to her opponent's, it provides an objective and publicly known measure of that ability. Second, the Elo rating allows us to control for the relative abilities of the two individuals competing, a feature often absent in studies of competitions where proxies for skill generally need to be used.

The players in our sample are exceptionally good at the game. A rating of 2000, the lower bound for players included in our sample, puts a player in the top 5 percent of all registered FIDE players. A rating of 2350, roughly the mean for all players in our sample, puts them in the top 0.5 percent of registered FIDE players. The mean Elo rating for male players is 2370.06 in our full sample (2342.98 for switchers), which is about 90 Elo points higher than the mean

[5] Note that the ratings of computer programs (there are two well known rating lists: SSDF list (Swedish Chess Computer Association) and the CCRL list (Computer Chess Rating Lists)) have only be computed through games with other computer programs and as such are an estimate. The question whether those ratings are directly comparable to the ratings of human players is debated among computer scientists and chess experts. However, although the 3200 Elo rating of the best computer programs is indeed an estimate, it is generally safe to assert they are much stronger than any human player.

[6] By FIDE rules, $K = 25$ if a player has played fewer than 30 games. Given the skill of the players in our sample, we assume all players have played more than 30 games.

rating of women. Women's opponents have lower average Elo ratings than men's opponents because players play opponents with similar Elo ratings as the majority of events have tiered entries, i.e. they follow the so-called Swiss system. We return to this point below.

The quality of play measure

In competitive activities, rating systems are generally accepted as a way to assess the relative skill levels of the participants. While there are numerous approaches to rating competitors (Elo ratings being just one example), these systems tend to be based on the realized outcomes of competitions.[7] Elo is a good measure of a player's overall ability but tells us nothing about how well a particular game is played. This distinction is important because we want to study whether the gender composition of a game affects how well a player plays a particular game.

Measuring within game play quality differs conceptually from studying strategic decisions. Others (e.g. Gerdes and Gränsmark, 2010) have sought to study the relationship between gender and strategic choices in chess focusing on the variation in the 'aggressiveness' of play, as defined by expert players. However, chess is ultimately a computational problem, which is precisely why computers excel at it.[8] The game of chess is theoretically solvable (Schwable and Walker, 1999).[9] That is, there is an optimal move for any given board position which can be calculated via backward induction. Any deviation from that move, be deemed 'aggressive' or otherwise by a human player, is suboptimal to some degree. How well a game is played can be objectively determined by the deviation of a move relative to the optimal move. Therefore, we depart from the literature that has considered more subjective, interpretive concepts such as 'style of play' or 'aggressive/solid move' (*how* a game is played) and instead consider the quality of play (how *well* a game is played).

To do so, we use a recently developed method by Guid and Bratko (2006, 2011) which allows us to assess the quality of the move played by each player for a given board position. The basis for this assessment is the difference between the move played by the human player

[7] For a comprehensive overview of such ratings in chess see Glickman (1995).

[8] As John von Neumann once noted 'Chess is not a game. Chess is a well defined form of computation.' (Bronowski, 1973).

[9] While the chess game is in principle solvable, it has never been solved. According to Shannon (1950) a typical game of 40 moves involves 10^{120} variations to be calculated from the first move. A computer calculating at the 'rate of one variation per micro-second would require over 10^{90} years to calculate the first move!'.

and the 'optimal' move as chosen by a powerful chess program.[10] For the current paper we use the powerful *Houdini 1.5a x64* program which has a maximum Elo rating of 3126, several hundred points above even the very best human players in history[11]. The Elo curve suggests this program would defeat the average player in our sample every time they played. Following Guid and Bratko (2006, 2011), we base our quality of play variable on the analysis of moves $n = 15, \ldots, 30$, in each game g with total length N moves. We then calculate 32 optimal moves (technically they are called plies), 16 for the player and 16 for the opponent. We consider this subset of moves for two reasons. First, to limit the substantial computational burden of calculating so many moves (about 1.5 million in our full sample). Second, we want to focus on the middle game, which is least likely to follow an established plan as expert players tend to study the opening moves of their opponents and practice end games in advance. For each of these moves, the chess program determines its preferred move given the position on the board with a search depth of 15 moves.[12] The chess engine effectively evaluates a decision tree that extends 15 moves forward from the move in question, evaluating the best move of both the player and the opponent at each node; this process encapsulates billions of possible board positions.[13]

We measure a player's relative advantage at a point in the game using the widely accepted metric called a centipawn. A centipawn is equal to 1/100 of a pawn.[14] A player in a given position with a score of 100 centipawns is seen as having an advantage equivalent to having an extra pawn on the board.

The quality of each move is measured by the difference between the centipawn (dis)advantage given the n^{th} move made by the chess engine, $C^n_{computer}$, and the centipawn (dis)advantage given

[10] We put optimal in inverted commas because the move determined by the powerful chess engine may not be the truly optimal move for a given board position since chess has not been solved yet.

[11] http://www.computerchess.org.uk/ccrl/4040/rating_list_all.html

[12] While greater search depth is feasible, it rapidly increases the computational burden. In general, increasing search depth by one move doubles the required computing time required.

[13] We include a non-technical discussion of how chess engines find optimal moves in Appendix A.

[14] For a given board configuration and a given chess engine configuration, the score in centipawns x can be interpreted as follows:

x		**Translates as**
x<-200	⇒	*Black* is winning
-200≤x<-50	⇒	*Black* is clearly better
-50≤x<-20	⇒	*Black* is slightly better
-20≤x <20	⇒	Approximately equal
20≤x<50	⇒	*White* is slightly better
50≤x<200	⇒	*White* is clearly better
200<x	⇒	*White* is winning

the n^{th} move actually made by the player, C^n_{player}, with larger differences indicating a larger error made by the player, i.e. a more poorly played move. We measure the quality of play for each player in each game, as the mean error committed by player i for moves $n = 15, \ldots, 30$

$$\overline{error_{ig}} = \frac{\sum_{n=15}^{\tilde{n}} \left(C^n_{computer} - C^n_{player} \right)}{\tilde{n}}, \tag{3}$$

where $\left(C^n_{computer} - C^n_{player} \right) \geq 0$ and $\tilde{n} = min\{30, N\}$ as some games end in fewer than 30 moves. We also calculate the mean error of the opponent, $\overline{error^o_{ig}}$. Note that larger values of $\overline{error_{ig}}$ indicate a lower quality played game.

Guid and Bratko (2011) considered dozens of games played by world champions, finding a mean error of about five. We consider tens of thousands of games and find a mean error of 16.5 (17.3 for opponents) in the full sample. As can be seen in Table 1, women commit larger mean errors, as might be expected given their lower average Elo ratings. There is a statistically significant negative correlation between a player's Elo and the mean error ($\rho = -0.18$, p-value$<$ 0.000) suggesting better players make smaller mean errors.

4 Analysis and results

Gender and performance

We first consider whether a player's gender plays a role in determining the outcomes of chess games. To do so, we regress the points earned from a game (one for a win, 0.5 for a draw, zero for a loss) on the player's gender -the effect of interest- while controlling for $E[p_{ig}]$ as defined in equation (2) to capture the difference in ability between the player and the opponent. We also include the age of each player and the color that i plays (white or black) as control variables. We estimate the model using OLS on pooled data and present results in Table 2.

In column (1), we use the full sample and in column (2) we use the sub-sample of switchers. Female players earn about 0.01 fewer points, on average and *ceteris paribus*, than their male counterparts. This simple result based on pooled data is consistent with much of the literature in this area which finds that women underperform men in competitive environments. This underperformance does not, however, seem to be simply a function of a player being female, nor can we deduce from this result that women are somehow innately worse players than men.

This is because there is another person, a male or female opponent, sat across from the player. In columns (3) and (4), we add a control for the gender of the opponent in game g. Results here suggest that it is not that players under or over-perform according to their own gender, but that it is the gender composition of the game what matters. Female players can expect to earn about 0.035 fewer points when they face a male opponent, even after controlling for the differences in their abilities via the Elo differential ($E[p_{ig}]$). Because of this, we next dwelve deeper in the study of the relationship between gender composition of games and performance.

GENDER COMPOSITION AND PERFORMANCE

In the analysis that follows, we explore the importance of the gender composition of games on the performance gender gap. To do so we exploit the panel dimension of our data and rely on within player i variation in the gender of the opponent. As players sometimes play a man and sometimes play a women, the opponent's gender might be conceived of as a 'treatment' which is applied in some games and not in others, the effect of which we want to study.

Any claim we make to the identification of the effect of this 'treatment' rests largely on the gender composition of a game being random, i.e. the genders of the player and the opponent being independent of one another. The advantage of laboratory experiments like Maas et al. (2008) is that they can explicitly randomize the gender composition of games. Because we study real competitions, we cannot randomly assign the gender composition of games, but we can still check whether the assignment is random or at least conditionally so.

The proportion of opponents who are female in our sub-sample of switchers is 0.22. The probability that a female player faces a female opponent is 0.61, much higher than the probability that a male player faces a female opponent (0.12). If the gender of the opponent were truly random, we would expect these values to match the proportion of opponents that are female. The fact that they do not indicates that the assignment of the opponent's gender, and thus the gender composition of the game, is not random.

The randomness of the gender composition of games is compromised by the presence of women-only events such as the Women's World Chess Championship,[15] and female-only sub-events taking place at larger mixed gender events. Female players can thus select out of playing

[15] While there are some tournaments that include only men by chance, there are no tournaments which exclude women as a matter of policy.

male opponents. Moreover, many tournaments have some form of seeding or tiers so that players of similar abilities end up playing each other. Given that men have higher average Elo ratings, this seeding system can further contribute to the correlation between the gender of the player and that of the opponent. The correlation between the Elo rating of the player and that of the opponent in our data is 0.449.

On the other hand, there is a random component in the assignment of players to games as the vast majority of tournaments employ a round robin format, with the players in the tournament playing one another once. Hence, players have virtually no control over who they end up playing with.[16] Given this random component in the assignment of players, the genders of the player and the opponent can be conditionally independent after controlling for both the gender composition of events and the mean Elo rating of players in the game.[17]

To test the conditional independence of the players' genders we regress the opponent's gender, a dummy equal to one if the opponent is male, on the player's gender, a dummy equal to one if the player is male, via OLS on pooled data. If the coefficient on the player's gender is not different from zero, it indicates that the genders of the player and the opponent are (conditionally) independent and that the gender composition of games is random. Results are presented in Table 3.

In column (1), the point estimate is 0.56 (95 percent CI: 0.537 to 0.588) indicating that a male player is 56 percentage points more likely to face a male opponent than a female opponent. In column (2), we re-estimate the model using the sample of 'switchers', i.e. only those players who play both men and women in our sample. For this sub-sample, the coefficient on the player being male falls to 0.49 (95 percent CI: 0.457 to 0.513). However, when we introduce the share of *other* players (excluding the player and opponent in game g) at the event who are female (column (3)) to capture the effect of women's tournaments and women only competitions taking place within larger tournaments, the coefficient on the player being male falls to 0.02 (95 percent CI: -0.000 to 0.041). In column (4), we include the mean Elo rating of the player and the opponent in game g ($\overline{Elo_g}$) to account for the fact that male players tend to have higher Elo ratings and players tend to play against opponents of similar ability. The point estimate of the coefficient on

[16] As chess tournaments often have a fairly large number of competitors, these round robin tournaments are generally of the Swiss-system variety where players play a pre-determined number of rounds, but fewer than a true round robin tournament.

[17] Given three variables x, y and z, the independence of x and y conditional on z requires that the conditional distribution of x given y and z, $p(x|y,z)$ does not depend on the value of y, so that $p(x|y,z) = p(x|z)$.

the player being male is effectively zero (95 percent CI: -0.020 to 0.020). In the last row of Table 3, we show the correlation and partial correlation coefficients which tell the same story. Namely, that once we condition on the mean Elo rating of the game and the gender composition of the event at which the game is played, there is no evidence of a relationship between the gender of the player and that of the opponent. We take this as evidence that, for the sub-sample of switchers, the 'treatment' in the form of the opponent's gender is conditionally independent of the player's own gender and that the gender composition of games is conditionally random.

We next exploit this conditional randomness to estimate the effect of the opponent's gender on a player's performance. To do so we estimate the following model:

$$points_{ig} = \alpha_i + \beta m^o_{ig} + \theta X_{ig} + e_{ig}, \tag{4}$$

where $points_{ig}$ is the points earned by the player i in game g, m^o_{ig} equals one if $i's$ opponent is male and zero if female, β, our parameter of interest, is the effect of the opponent's gender on the outcome of the game in terms of the points earned by the player, X is a vector of controls detailed below, α_i is a player i fixed effect capturing time invariant individual characteristics such as innate ability and preferences for an opponent's gender that may be correlated with X_{ig} and/or m^o_{ig}, and e_{ig} is a random error term with mean zero that is assumed to be uncorrelated with X_{ig} and m^o_{ig}, though we return to this issue below.

We estimate equation (4) using OLS on within-i mean differenced data to eliminate α_i. Our main results are presented in Table 4.

In column (1), we regress $points_{ig}$ on m^o_{ig} only. The coefficient of -0.10 (95 percent CI: -0.12 to -0.09) indicates that the player earns on average 0.1 percent fewer points when the opponent is male. In column (2), we add the control vector X: the ages of the player and the opponent, the player's expected points as calculated in equation (2), dummies for the opponent's affiliated national chess federation, the color being played by the player, the mean Elo of the player and the opponent, and the share of *other* players at the event who are female. The point estimate of $\hat{\beta}$ falls in absolute value to -0.04 (95 percent CI: -0.053 to -0.027). In column (3), we estimate the model using the restricted sample in which the player plays both genders at least once ('switchers'). As discussed above, it is for this sub-sample that the gender composition of the game is conditionally random. The estimated $\hat{\beta}$ remains -0.04 (95 percent CI: -0.054 to -0.027).

We then allow the effect of the opponent's gender to differ for male and female players since

so far we have been mixing two types of games in the reference category: male player vs female opponent and female player vs female opponent. To address this, we split the sample according to players' gender and re-estimate equation (4). In column (4), we estimate the model using only female players who have played both men and women in our sample. The magnitude of the point estimate maintains (95 percent CI: -0.080 to -0.007), i.e. female players earn fewer points against male opponents. We estimate the model for only male players who have played both men and women in column (5) and find a very similar result (95 percent CI: -0.057 to -0.028). A t-test of the equality of the $\hat{\beta}$ coefficients in columns (4) and (5) returns a p-value of 0.961.[18] We thus find no evidence that the effect of the opponent's gender differs with the gender of the player.

These results indicate that players earn, on average and *ceteris paribus*, about 0.04 fewer points when playing against a man as compared to when their opponent is a woman. Or conversely, men earn 0.04 points more when facing a female opponent than when facing a male opponent. This is a sizable effect, comparable to women playing with a 30 Elo point handicap when facing male opponents. Such an effect indicates some change in behaviour when people engage in inter-gender competition. What we cannot say from simply looking at the effect of inter-gender competition on outcomes is whether it is the behaviour of men, women or both which changes. We study this below, but first we test the robustness of our main results from Table 4.

Supplementary analysis

In this section, we carry out supplementary analyses to test the credibility of the identification strategy and the results above. First, we test the robustness of the primary results to mis-specification. Second, we consider a particular mis-specification in the form of a possibly unaccounted for non-linearity in the relationship between the outcome of the game and the Elo differential, as controlled for by $E[p_{ig}]$. Such a non-linearity might in turn be correlated with the gender of the opponent and thus bias our results. Third, we consider the bias arising from the measurement error in Elo ratings that a genuine gender effect, i.e. $\beta \neq 0$, would introduce.

[18] We estimated a fully interacted version of equation (4). This p-value is from the t-test of the coefficient on the interaction of the player's and opponent's genders being equal to zero.

Robustness to misspecification

To test the robustness of our results to mis-specification we follow the good practice outlined in Athey and Imbens (2015) and re-estimate equation (4), via OLS using within-i mean differenced data for different sub-samples. The results are presented in Table 5. In column (1), we estimate the model excluding any games played in single sex tournaments, either those explicitly women only, or those accidentally all male or all female. In column (2), we follow Gerdes and Gränsmark (2010) by estimating the model excluding games that ended in a draw, and in column (3) we estimate the model using players who play at least 20 games in our sample. In column (4), we exclude blitz events and junior events. Lastly, in column (5), we replace the gender composition of the event variable with event fixed effects. In each case, the magnitude and statistical significance of the effect of the opponent's gender maintains.

In Table 6, we re-estimate equation (4) for different sub-samples defined by the Elo differential and the Elo rating of the player. In column (1), we include only games where the Elo differential between the player and the opponent is less than or equal to 200 Elo points; in column (2) we use only games where the Elo differential is less than or equal to 100 Elo points; and in column (3) less than or equal to 50 Elo points. In column (4), we exclude games played by players with Elo ratings less than 2200, and in column (5) we exclude those with Elo ratings less than 2400. Again, in each case, the magnitude and statistical significance of the result maintains.

Non-linearity in the $E[p_{ig}]$-outcome relationship

As discussed above, women have, on average, lower Elo ratings than men, though they also face opponents with lower average Elo ratings. However, when a female player in our sample plays a male opponent, she faces an average disadvantage of 27 Elo points as opposed to a mean advantage of 23 Elo points when she faces a female opponent. The Elo differential is thus correlated with the gender of the opponent. The correlation between the opponent being male and the $E[p_{ig}]$ is small but significant ($\rho = -0.049$, p-value< 0.000). We control for the Elo differential via $E[p_{ig}]$ and for the mean Elo of the game with $\overline{Elo_g}$. Still, we may be neglecting some non-linearity in the effect of the Elo differential or of $\overline{Elo_g}$ on the outcome of games. We test the robustness of our results to more general specifications of the Elo ratings and the Elo differential using switchers only, and present the results in Table 7.

In column (1), we replace $E[p_{ig}]$ and $\overline{Elo_g}$ in equation (4) with the logged Elo ratings of

the player and the opponent. In column (2), we replace $E[p_{ig}]$ and $\overline{Elo_g}$ in equation (4) with dummies for decile groups of $\overline{Elo_g}$ and $E[p_{ig}]$. In column (3), we add the squares an cubes of $\overline{Elo_g}$ and $E[p_{ig}]$ to equation (4). In column (4), we include the interactions of the player and the opponent's logged Elo ratings with the decile groups of $\overline{Elo_g}$ and $E[p_{ig}]$, allowing the effect of Elo ratings to vary depending on the relative (dis)advantage of the player and the average ability of the player and opponent in the game. In column (5), we add a dummy equal to one if the player is at an Elo point disadvantage and zero otherwise to our baseline specification.

It is encouraging that the estimated gender effect of the opponent being male remains markedly stable in both magnitude and precision as the specification of the effect of Elo ratings becomes increasingly flexible.

Measurement error bias

Next, we address the issue of a potential measurement error in Elo ratings resulting from the effect of the opponent's gender which may bias the estimator of β. Under the null hypothesis that $\beta = 0$, the Elo rating is a reliable measure of players ability. However, under the alternative that $\beta \neq 0$, Elo ratings would systematically measure with error the ability of any player who plays members of the opposite sex, given that women (men) under- (over)-perform when playing against men (women). That is, if the gender of the opponent matters, Elo ratings would systematically mis-measure the true ability of players compared to the case where the gender effect is absent or where players are unaware of the gender of their opponent.[19] In particular, if women perform worse against men, even after controlling for Elo differentials, women's Elo ratings would systematically under-rate their true ability whereas men's Elo ratings would systematically over-rate their ability. Players' Elo ratings would then tend to be 'too big' for men and 'too small' for women. This measurement error in the opponent's Elo rating would therefore be correlated with m^o_{ig} and would bias the OLS estimator of β. It is important to keep in mind that this measurement error, and thus the resulting bias, would be present if and only if $\beta \neq 0$. Therefore, this measurement error cannot be responsible for the significance of our result but it may lead us to underestimate its magnitude.

To see this formally, consider a simplified version of our model

[19] This mis-measurement would also affect players who only play opponents of their own gender but whose opponents' opponents have been of the opposite sex and so on.

$$points_{ig} = \beta m^o_{ig} + \theta x^o_{ig} + \eta_{ig}, \tag{5}$$

where $points_{ig}$ is the points earned by player i, x^o_{ig} is the measured Elo of the opponent, m^o_{ig} is the dummy equal to one if the opponent is male and η_{ig} is an error term uncorrelated with x_{ig} or m^o_{ig}. Our problem is that x^o will be measured with error such that $x^o = x^{o*} + w$ where x^{o*} is the true Elo rating of the opponent and w is positively correlated with m^o_{ig} if $\beta \neq 0$.

To derive the bias in our estimator of β we first regress m^o_{ig} on x^o_{ig} and obtain the residuals, ι. We then regress $points_{ig}$ on x^o_{ig} and obtain the residuals, τ. Finally, we regress τ on ι to obtain

$$\hat{\beta} = (m'Pm)^{-1}(m'Py), \tag{6}$$

where $P = I - x(x'x)^{-1}x'$ is the symmetric idempotent matrix with $Px = 0$. Restating equation (6) gives

$$\begin{aligned} \hat{\beta} &= (m'Pm)^{-1}(m'P(\beta m + \theta x + \eta - \theta w)) \\ &= \beta + (m'Pm)^{-1} m'P(\eta - \theta w). \end{aligned} \tag{7}$$

Taking expectation throughout yields

$$E\left(\hat{\beta}\right) = \beta + E\left[(m'Pm)^{-1} m'P\eta\right] - \theta E\left[(m'Pm)^{-1} m'Pw\right]. \tag{8}$$

The second term in equation (8) equals zero as $E[\eta m] = 0$ by assumption. The third term depends on the covariance of m and w, which is positive, and on θ, which is negative (the points i can expect to earn in a game decrease with the Elo of the opponent). Therefore, the bias in $\hat{\beta}$ is positive, i.e. towards 0. This means that our estimates of β in Table 4 can be interpreted as lower bounds (in absolute value) of the true effect of the opponent's gender on i's performance. That is, the true gender effect may be larger than what we find.

The key result here is that the bias cannot drive our finding that performance is diminished by playing against a male opponent since *i*) the bias is only present if there is indeed a gender effect; and *ii*) the bias is towards zero, meaning that we are underestimating, in absolute value, the true gender effect.

5 Why do women underperform against men in chess?

So far we have established that players fare worse when their opponent is male. However, we cannot say whether this is due to changes in men's behaviour, women's behaviour or both. Moreover, the mechanism through which this effect operates still remains unclear. In this section, we first discuss three popular explanations for the underperformance of women in chess and argue that they fail to account for the observed effect. We then explore three mechanisms consistent with the observed gender effect: variation in the quality of play, variation in the competitiveness of players, and variation in the response to competitive pressure.

Innate ability, practice and strategy

One of the most popular explanations for the gender gap in the performance of expert players refers to innate gender differences in cognitive abilities key in chess. As the opening quotes demonstrate, the perception that men are superior players persists. This view is best exemplified by Howard (2004), who uses the substantial and persistent gender differences in Elo ratings to conclude that men are inherently superior to women in this domain. However, there is no compelling evidence that either men or women are better suited to excel at chess. Some have contended that men's better performance on spatial rotation tasks provides them an advantage (Li, 2014). Other evidence suggests that recognizing positions on chess boards is more akin to recognizing faces (Boggan, Bartlett and Krawczyk, 2012), a task at which women outperform men (Herlitz and Lovén, 2013). A key piece of evidence in this debate comes from Bilalic et al. (2009) who show that the observed superiority of men in chess, as measured by the average Elo rating of men versus women, is almost entirely due to the characteristics of extreme value distributions. Once sampling is taken into account, there is virtually no gender difference in mean Elo ratings. But even if in spite of the evidence, one were to believe men to be innately superior chess players, that would not explain why women perform worse when playing against a man, all else, including Elo, being equal.

Another potential explanation for the observed gender effect could be differences in deliberate practice and dedication. Chess expertise requires intense, almost obsessive, training and constant practice and study, which can in turn interfere with child-rearing. Leaving aside the hours of coaching they receive, top chess players accumulate on average 5,000 hours of study alone by the tenth year of their career. This figure is comparable to the level of deliberate practice

accumulated by symphony-level musicians (Charnes and Gerchak, 1996). De Bruin, Smits, Rikers and Schmidt (2008) find that differences in Elo ratings are partially explained by gender differences in the time devoted to deliberate practice (though they fail to account for the sampling issue raised in Bilalic et al. (2009)). However, they also find that gender differences in ratings remain significant after controlling for the amount of deliberate practice. Therefore, differentials in practice and study cannot account either for the gender performance gap between equally skilled male and female players. And again, even if men were more dedicated chess players, that would not explain why women perform worse against men, all else, including Elo, being equal.

A final possibility is that mixed-gender games may result in players making different strategic choices, which result in worse outcomes for women. As discussed above, this possibility is considered in Gerdes and Gränsmark (2010) and Maas et al. (2008). Gerdes and Gränsmark (2010) find that men choose more 'aggressive' opening strategies when playing against women. They find, however, that these more 'aggressive' strategies actually reduce the odds of men winning and thus cannot explain the performance gender-gap we observe. Shahade (2005) suggests that the style of play of women in the Elo range of 2300-2500 is commonly viewed as excessively aggressive and impatient. Maas et al. (2008) find however that women are less likely to declare aggressive 'intent' at the start of games played against a man. They also find, in contrast to Gerdes and Gränsmark (2010), that men's declared aggressiveness is not affected by the gender of the opponent. Unfortunately, Maas et al. (2008) do not relate the declared aggressiveness of players to outcomes nor use Elo ratings to control for the ability of players.

In summary, the current evidence only considers strategic variations in a single dimension, 'aggressiveness,' as determined by the subjective interpretation of play style. Ultimately, this approach fails to explain our results. For these reasons, we depart from the approach of studying play style and instead focus on the more objectively measurable quality of play. We use the measure of within game play quality presented above to examine whether the gender of one's opponent affects how well a player is able to play.

The effect of inter-gender competition on Quality of play

Although we have already discussed our quality of play measure in detail, let us reiterate that the key element of the metric is that, unlike the Elo rating, it is not a measure of a player's overall ability or skill. Rather, it captures how well a particular game was played based on the

analysis of individual moves. This allows us to determine how one's capacity to play chess, not just the observed outcomes, varies with the gender of one's opponent.

While we find that women make larger errors on average than men (see Table 1), we do not find evidence that women make larger errors than men of equivalent Elo rating. We find that there is no difference between the mean error ($\overline{error_g}$) of a female-female game and that of a male-male game once we condition on the mean Elo rating of the game ($\overline{Elo_g}$).[20] Moreover, the Elo elasticity of errors is the same for male and female players indicating that the relationship between ability and quality of play is not gender specific.[21] Our conjecture is that the gender difference in errors is driven by the gender composition of the game, all else being equal.

To test the impact of gender composition on the quality of play, we estimate the following model

$$ln\left(\overline{error}_{ig}\right) = \delta_i + \gamma m^o_{ig} + \theta X_{ig} + u_{ig}, \tag{9}$$

where the dependent variable is the logged mean error committed by the player in game g and u_{ig} is a random error term. The model is estimated via OLS on within-i mean differenced data to eliminate δ_i using the same set of controls as in equation (4). Results are presented in Table 8. In columns (1) and (2), we report the results for female players (full and switchers samples, respectively) and for male players in columns (3) and (4) (full and switchers samples, respectively).

We find that the mean error committed by a female player between moves 15 and 30 increases by about eight percent when facing a male opponent (95 percent CI: 0.006 to 0.168 in column (1) and -0.000 to 0.168 in column (2)). This indicates that female players are playing worse, on average and *ceteris paribus*, when their opponent is male. We do not find evidence that the quality of a male player's play is affected by the gender of the opponent. Columns (3) and (4) show point estimates very close to zero. We test the equality of the effect of the opponent's

[20] We regress $\overline{error_g}$ on $\overline{Elo_g}$ and dummies for male-male games and for female-female games (the base group being inter-gender games) and then test the equivalence of the estimated coefficients on those dummies obtaining a $p-$value of 0.642. Thus we do not reject the null that $\overline{error_g}$ is equal in male-male and female-female games when we condition on $\overline{Elo_g}$.

[21] We regress the logged mean error of the player on the log of the player's Elo rating and controls for the logged Elo rating as well as all the controls on our main regression all interacted with the gender of the player. We estimate this model via OLS on within-i mean differenced data. The $p-$value of the estimated coefficient on the interaction of the player's gender and the player's logged Elo rating is 0.580. Thus we do not reject the null that the Elo rating elasticity of errors is the same for male and female players.

gender on the quality of play of male and female players. Results suggest that the effect does differ for male and female players (p-value=0.026 for the switchers sample).[22]

The quality of play does explain some of the observed gender effect in Table 4. When we include the log error of the player and re-estimate equation (4) using the sample of female switchers (equivalent to column (3) of Table 4) we find the size of the estimate ($\hat{\beta}$=-0.032, 95 percent CI: -0.044 to -0.019) somewhat reduced relative to the baseline. This reduction is statistically significant (p-value=0.006). We subject the analysis of play quality to the same robustness checks as our primary analysis and report results of these checks in Appendix B where we find strong support for our result here: female player's quality of play is reduced by about eigth percent when facing a male opponent, and a male player's quality of play is not affected by the gender of the opponent.

Stereotype threat

Let us now consider three mechanisms which could explain the effect of male opponents on female's quality of play.

The first one is stereotype threat, which occurs when individuals perform worse in a task in fear of confirming a negative stereotype applying to the group they belong to (Steele, 1997). The anxiety, self-doubt and negative feelings that belonging to a stigmatized group generates undermine performance, thus confirming the negative stereotype. Salience of negative stereotypes has been shown to lead to higher heart-rate variability paired with poorer performance in tests (Croizet, Després, Gauzins, Huguet et al., 2004), and to higher activation of brain areas related to emotions (Krendl, Richeson, Welley and Heatherton, 2008). Stereotype threat is thus a natural candidate to explain the observed gender effect on play quality.

Negative stereotypes about female players are very prevalent and salient in chess, as the opening quotations demonstrate. According to stereotype threat theory, these negative stereotypes may disrupt the cognitive capacity of females playing against males, thus reducing their quality of play. Chess requires computational effort to determine the optimal move at each stage of the game. Given a level of innate ability, better outcomes require deeper computation which

[22] We estimated a fully interacted version of equation (9). This p-value is from the t-test of the coefficient on the interaction of the player's and opponent's genders being equal to zero.

in turn requires greater mental effort.[23] As such, any disruption in the ability to commit mental resources to the computation of the best move can manifest itself in lower quality of play.

Stereotype threat has been shown to impair a cognitive resource key in chess: working memory. Working memory is defined as the ability to focus the attention on a task and to store task relevant information temporarily, whilst inhibiting task-irrelevant information and thoughts. Robbins, Anderson, Barker, Bradley et al. (1996) introduced secondary tasks aimed at interfering with working memory and found a detrimental effect on chess performance. This suggests that if female players feel stereotype threat when competing against a male player, stress, excessive self-monitoring and the need to suppress negative thoughts and regulate anxiety might tax their working memory and harm their performance in the game. This should naturally be reflected in our quality of play measure, as Table 8 shows.[24]

The quality of play metric allows us a deeper insight. In particular, it allows us to establish whether the gender effect described in Table 4 is consistent with stereotype threat, that is, women playing worse when they play men, or with stereotype boost (or lift), i.e. men playing better when they play women (Walton and Cohen, 2003).[25] Results in Table 8 are consistent with stereotype threat but not with stereotype boost.

We cannot however, take the results in Table 8 as conclusive evidence of stereotype threat at work. The evidence provided here is consistent with the impairment of cognitive abilities that stereotype threat theory predicts in negatively stereotyped groups, females in the case of chess.

Competitiveness

The second explanation for the reduction in females' quality of play we observe is a decrease in the effort or willingness to compete of female players when facing a male opponent. This may arise from under-confidence (Niederle and Vesterlund, 2011) or from distaste for competition,

[23] Dr. Emanuel Lasker, a German chess player, mathematician, and philosopher who was World Chess Champion for 27 years is quoted as saying 'When you see a good move wait - look for a better one.'.

[24] Another channel by which stereotype threat can reduce chess performance is through a disruption of spatial abilities. Wraga, Helt, Jacobs and Sullivan, (2006) used fMRI to show that stereotype threatened females perform worse in a mental rotation task, and that this poorer performance is due to a higher activation in brain regions associated with emotional loads. These results might be relevant for chess because it is a visually-demanding activity. Players should keep track of the positions of a great number of pieces and need to learn a large number of piece combinations.

[25] Several studies show that performance can increase among the positively stereotyped group at the same time as performance decreases for the negatively stereotyped group (Wraga et al., 2006).

especially against males (Croson and Gneezy, 2009).[26] Reduction of effort is consistent with the lower quality of play that female players exhibit when playing against males. Although we do not observe effort directly, it is possible to study competitiveness indirectly by looking at resignations and the length of games.

The vast majority of losses in expert chess are due to resignations, where a player concedes defeat before technically losing, with only 1.2 percent of games in our sample being played through to checkmate. These resignations could be interpreted as a reluctance to engage in further competitive effort once the player considers that the expected outcome of the game is not good enough to warrant the required effort to continue playing. When deciding whether to resign or not, players perform a cost-benefit analysis. Continuing the game at a given board position has the cost of the additional effort exerted and the benefit of the additional expected points earned. If females are under-confident or have a distaste for playing men, we would expect them to resign more quickly, i.e. in fewer moves, when playing against a male opponent, all else being equal.

Our interest then is whether the gender of the opponent plays a role in determining the length of a game. The average game in our sample lasts 43 moves. In average, the longest games are female-female games with an average number of moves of 45.9. Male-male games are shorter, only 42.6 moves on average. Mixed gender games fall in between at 43.3 moves per game on average. To formally test these differences we identify those games in which player i resigns and we calculate the total number of moves. We then regress the logged number of moves in game g where the player has resigned on the same set of controls as in equation (4) plus the logged quality of play of both the player and the opponent ($ln\left(\overline{error}_{ig}\right)$ and $ln\left(\overline{error}^{o}_{ig}\right)$). Results are presented in Table 9. In columns (1) and (2), we report the results for female players (full and switchers samples, respectively) and for male players in columns (3) and (4) (full and switchers samples, respectively).

We do not find compelling evidence that the gender of the opponent affects female players' readiness to resign, i.e. no reduction in effort or willingness to compete. The point estimates in columns (1) and (2) are negative, but the confidence intervals are too wide for meaningful inference (95 percent CI: -0.114 to 0.037 in column (1) and -0.112 to 0.045 in column (2)). The

[26] Using survey data, Kleinjans (2009) finds that, controlling for ability and family background, the median female expresses greater distaste for competition than the median male. Female's stronger distaste for competition lowers their educational attainment relative to that of male.

evidence suggests that men, however, resign in about eight percent fewer moves against a male opponent (95 percent CI: -0.115 to -0.032 in column (3) and -0.122 to -0.038 in column (4)), though we cannot reject the equality of this effect for male and female players (p-value=0.297).[27] Men resign in fewer moves against other men. This suggests additional explanation for the observed gender effect: an extra willingness of men to compete against women.

One possibility is that men know that women are more error prone when playing male opponents. As a result, males may play longer against a female opponent, holding out for the woman to make a larger error than a male opponent would. Were this the case, however, we should also expect that women, knowing that they make larger errors against men, resign more quickly against a male opponent. But we find no evidence of this.

An alternative explanation might be that the increased competitiveness of men stems from a psychological cost to men of losing to a woman. In the case of chess, anecdotal evidence suggests that such cost may be very real.[28] If that is the case, given two identical board positions, and two opponents of the same ability and who are playing equally well, a male player will be more likely to continue playing against a female opponent than against a male opponent.[29] By persisting longer in a disadvantaged situation, i.e. 'dragging it out', a male player facing a female opponent maintains a non-zero probability of forcing a draw or even winning the game thus earning, on average, more points against a female opponent than he would against a male opponent all else equal. This, of course, means that female players in these games earn, on average, fewer points as a result.

The two mechanisms described above suggest that the gender composition of a competition affects the behavior of both male and female competitors. Women tend to play worse when facing a male opponent, all else being equal. Men are less willing to yield when facing a fe-

[27] We estimated a fully interacted version of the model. The p-value reported here is from the t-test of the coefficient on the interaction of the player's and opponent's genders being equal to zero.

[28] American essayist Charles Dudley Warner famously quoted that 'There is nothing that disgusts a man like getting beaten at chess by a woman.'. Much more recently, in the thread 'Do men dislike losing to women, if so why?' in the Chess.com forum, a user writes: 'I've found male players will drag it out to the last minute, even when it's clear they should resign, or are in check or about to be mated, they will still wait one day or three days before moving, why is this, it's so annoying.'.

[29] Consider a simple illustration of this mechanism. For a male player, the payoff from resigning to a man is zero. If the player decides to continue the game, his expected payoff in terms of the outcome of the game is $E[p|pos^n] - e$, where e is the effort of continuing the game, and $E[p|pos^n] \geq 0$ is his expected points given the board position at move n of the game. A male player will resign against a male opponent if $e > E[p|pos^n]$. Now suppose the payoff to the same man of resigning to a woman is $-c$, where c is a psychological cost of losing to a woman. A male player will then resign against a female opponent if $e - c > E[p|pos^n]$. In other words, *ceteris paribus*, the $E[p|pos^n]$ required for a male player to resign against a woman is lower than the $E[p|pos^n]$ required for him to resign against a man.

male opponent. While the two mechanism are very different, they both serve to diminish the performance of female competitors.

Do the stakes of the game matter?

Let us next explore a third mechanism explaining the effect of the gender composition of games on females' quality of play. This explanation is based on the emerging evidence showing that gender differences in competitive performance might be due to differential responses to competitive pressure. For example, using data from Grand Slam tennis tournaments, Paserman (2010) finds that both men and women show a substantial deterioration in performance as points become more important, and that women, not men, increase the ratio of unforced errors to winning shots as stakes increase. Jurajda and Munich (2011) and Örs et al (2013) study the effect of stakes on performance in the context of college admissions. Jurajda and Münich (2011) use data from different university entry exams taken by the same individuals and find that women underperform men in the access to top universities but not elsewhere. Similarly, Örs et al. (2013) find that females perform worse than males in the entrance exam for the very selective HEC Paris despite having performed better in the less competitive Baccalaureat exam.

If, as this literature suggests, the competitive performance of women deteriorates with competitive pressure more than men's, and the competitive pressure in mix gender games is higher than in single gender games, the combination of these two effects could be responsible for the lower quality of play of women when playing against men. We next explore this possibility.

How to measure stakes in chess? Most tournaments have a prize fund from which prizes are awarded. In some tournaments, there are special prizes for younger or females players. Variability in monetary stakes is thus large. Given that most tournaments are round robin rather than elimination competitions, one would need very detailed information on the order in which a player faces his or her opponents in order to asses accurately the monetary relevance of each single game; unfortunately, this information is generally not available.

Expert chess games have another type of stakes: the potential gain or loss in Elo rating. Changes in Elo ratings are important for both economic and psychological reasons. The economic reason is that organizers and sponsors of competitions use Elo ratings to chose the field

of participants.[30] Highly-rated players can also demand substantial appearance fees. The psychological importance of ratings has to do with status. The Elo rating provides an absolute measure of ability that players can use to asses their own quality and to measure themselves against others. Traditionally, FIDE updated the world ranking of players twice a year. Today, a number of websites provide rankings updated in real time.[31] Many threads in chess forums contain debates and discussions about various approaches to increasing one's Elo rating, and on whether the much prevalent obsession with ratings in expert chess is counterproductive for one's game.[32] Hence, a player's expected gain or loss of Elo points is a good proxy for his/her stakes in a particular game, and it is therefore a relatively good measure of competitive pressure.

We study whether the Elo points at stake in a particular game affect the quality of play. To calculate the Elo points at stake we follow equation (2) in determining the expected points of the game where, again, a win is worth one point (*not* Elo points), a draw worth 0.5 and a loss worth zero. This does not, however, allow us to directly calculate the expectd Elo points at stake in the game. The change in the player's Elo from equation (1) is $K\left[p_{ig} - E[p_{ig}]\right]$, but this depends on the actual outcome of the game, p_{ig}. A player at an Elo disadvantage will gain more Elo points from winning the game than she will lose from losing the game. The same player will gain points from a draw, while a player with an Elo advantage will lose points from a draw with a lower rated player.

Therefore, to calculate the expected Elo points from a game, we need to estimate the probability distribution over outcomes. The stakes of the game will be based on the expectation that the player will win (π_w), lose (π_l) or draw (π_d). The estimates of these outcome probabilities must be empirical rather than analytical, i.e. cannot be based on $E[p_{ig}]$.[33]

We estimate π_w, π_l and π_d non-parametrically using a large dataset of about 1.5 million

[30] Elo ratings play an important role in team championships. Many professional and semi-professional players earn a substantial part of their salaries as players on teams in national leagues and invitations to join such teams are based largely on Elo rating. The same applies to national teams competing internationally where only few players (there are two national teams: male and female) are invited.

[31] See http://www.2700chess.com/ for one example.

[32] See, for instance, 'Obsessed with ratings' in Chess.com https://www.chess.com/forum/view/general/obsessed-with-ratings.

[33] The Elo curve described in equation (2) is the way in which FIDE computes the expected points of a player in a game. Hence, it does not need to coincide with the empirical expectation. If they did, players' expected gain/loss from a game would be zero.

games played between 1900 and 2008.[34] We use this larger, secondary data source so that we can produce highly accurate estimates of π_w, π_l and π_d. We need this because the actual probability distribution over these outcomes varies with the level of Elo. For example, the probability of a game played between players in the bottom decile of our Elo distribution ending in a draw is 0.25, whereas the probability of a game played by players in the top decile ending in a draw is nearly double, 0.48. To account for this, we construct mean Elo rating $\left(\overline{Elo_g}\right)$ percentile groups and also Elo difference $\left(Elo_{ig} - Elo^o_{ig}\right)$ percentile groups. We then calculate π_w, π_l and π_d for the intersections of each of these, giving us 1000 estimates of these probabilities over both mean Elo rating and Elo difference. Using these probabilities we construct a new variable, *stakes*, defined as

$$stakes_{ig} = \left|K(0.5\hat{\pi}_{ig,d} + \hat{\pi}_{ig,w} - E[p_{ig}])\right|, \tag{10}$$

which is the absolute value of the expected Elo points to be earned or lost by player i in game g and $\hat{\pi}_{ig,w}$ and $\hat{\pi}_{ig,d}$ are the estimated probabilities of winning and drawing for the Elo differential-mean Elo percentile group intersection in which game g is located. This *stakes* variable is our measure of the stakes of the game. On average, there are just over two Elo points at stake, though stakes in inter-gender games are slightly higher (2.2 for males and 2.4 for females) than in single gender games (1.9 for female-female and 1.7 for male-male). Hence, if as we have conjectured, females' competitive performance deteriorates more than males' as competitive pressure increases, the increased stakes in mix gender games could contribute to explain the detrimental effect of gender composition on the women's quality of play.

It is the impact of stakes on the quality of play that we are interested in. To examine this we include the log of the *stakes* variable into our model of quality of play in equation (9) and re-estimate the model via OLS with player i mean differenced data using switchers only. Results are presented in Table 10. In column (1), we present the result for female players. Larger stakes do increase the mean error committed by female players, though the effect seems small: a ten percent increase in the stakes leads to a 0.34 percent increase in the mean error. This is consistent with results in Paserman (2010). In column (2), we interact log stakes with the gender of i's opponent for female players and report the marginal effects of the logged stakes from this

[34] These are from a database of PGN files stored at http://www.top-5000.nl/pgn.htm. No FIDE registration numbers are available in this data set, so we could not carry out the above analysis with it as we cannot identify the gender of players.

interaction. While the point estimate of the coefficient on the interaction is positive, indicating that the effect of the stakes on play quality may be larger when the opponent is male, it is not statistically significant (p-value=0.292) meaning that we cannot draw a conclusion about whether the effect of the stakes of the game vary with the gender of the opponent.

The stakes also affect the quality of play of male players (column (3)) with a ten percent increase in the stakes leading to a 0.3 percent decrease in quality of play. Though the point estimate for the impact of stakes on play quality is smaller for men than for women, the difference is not statistically significant (p-value=0.385). In column (4), we again interact the logged stakes with the gender of the opponent. We do not find statistical evidence that the effect of the stakes of the game for male players depends on the gender of the opponent (p-value=0.615). These results suggest that competitive pressure does have a detrimental effect on play quality, though we are unable to convincingly detect any heterogeneity in that effect over the gender composition of games.

6 Conclusions

In this paper, we study gender differences in competitive performance in a real world setting. We use data from tens of thousands of chess games played by thousands of expert players. Like women in high-power jobs, expert female chess players have selected into a male dominated and highly competitive arena. This sample selection is to our advantage. We study a population that is less likely to experience gender differences in competitiveness than a randomly selected sample. Although, as in the laboratory, this may come at the cost of external validity, we identify gender effects in a real world setting where selection migth mitigate these effects.

We test for the presence of a performance gender-gap, first by looking at the performance of players by gender. Our results are consistent with previous literature showing that females perform worse than males of comparable ability in competitive settings. Then we show that this underperformance is driven by the gender composition of games. Our results show that women earn about 0.04 fewer points when their opponent is male, even when we control for player fixed effects, the ages and the skill levels (as measured by the Elo rating) of the players involved.

We then study the mechanism underlying the observed results. In doing this, we choose not to consider changes in strategic choices. We have doubts about whether it is at all possible to identify strategic variation in a meaningful way by looking at the early phases of the game. In

the famous 'Game of the Century' played in 1956 between a 13 year-old Bobby Fischer and Donald Byrne, Fischer shockingly lost his queen early on, a move that was seen as a mistake. Only later in the game did it become clear that the sacrifice was a brilliant piece of strategy by Fischer that led almost inexorably to his victory. Systematically codifying and rating moves in a game with so much variation, and where a player's strategy may not become clear until much later, is a formidable task and one we believe cannot be done, at least in the framework of the current paper.

Instead, we employ a unique measure of within-game quality of play developed in Guid and Bratko (2006, 2011). Unlike the Elo rating, which is a function of all previous games, this measure of quality of play compares the moves actually played within the game to the moves a powerful chess engine would have made in the same position. The distance between these two moves is a measure of the error committed by the player. Moreover, this measure focuses on the middle-game where, unlike for openings, creativity and improvisation are key. We study whether this error is a function of the gender composition of the game. Doing so allows us to test whether the gender effect is the product of women playing worse when they play against men or of men playing better when they play against women.

The results show that the mean error committed by women is about eight percent larger when they play against a male. This is likely to be a conservative estimate of the actual effect. This suggests that the variation in the quality of play explains some portion of the gender effect on outcomes. We interpret this result in the light of the stereotype threat literature. The differential response of errors to the gender of the opponent may operate via a reduction in the cognitive ability of female players facing male opponents. There is no evidence of stereotype boost affecting men. We do find, however, that men resign more quickly (after fewer moves) against other men than they do against women. It seems that men continue playing against women even when they would resign were they playing against men. This persistence, this willingness to continue competing against female opponents, is also consistent with the observed gender effect.

We finally consider the effect of competitive pressure as measured by the Elo points at stake in the game. This is unlikely to be the only source of variation in competitive pressure affecting players in a particular tournament, but we do not observe tournament round or prize money at stake, other potential sources of such pressure. We find evidence that an increase in the degree

of competitive pressure as measured by the Elo point stakes of the game undermines play quality of both men and women to a similar degree. Future research might consider whether this effect might depend on the gender composition of competitions, though we do not find persuasive evidence for such heterogeneity here.

We believe we have provided compelling evidence of the presence of gender effects on competitiveness in a sample of expert chess players. Our results suggest that the introduction of blind competitions at high level chess tournaments might be a desirable intervention. Blind auditions have shown to have a very positive effect on the representation of women in top orchestras (Goldin and Rouse, 2000). Perhaps more important is the fact that women in our sample have achieved a level of mastery in chess. If the effects we observe are present for such a selected sample, it seems reasonable to assume that they will be operating at least as strongly in a more general population. Further research should look into what approaches might be employed to mitigate this effect.

References

Antonovics, K., Arcidiacono P., and Walsh, R. (2009). "The effects of gender interactions in the lab and in the field", *Review of Economics and Statistics*, 91, 152–622.

Athey, S., and Imbens, G. (2015). "A measure of robustness to misspecification", *American Economic Review*, 105, 476-80.

Auster, E., and Prasad, A. (2016). "Why Do Women Still Not Make It to the Top? Dominant Organizational Ideologies and Biases by Promotion Committees Limit Opportunities to Destination Positions", *Sex Roles*, 75(5), 177-196

Bilalic, M., Smallbone, K., McLeod, P., and Gobet, F. (2009). "Why are (the best) women so good at chess? Participation rates and gender differences in intellectual domains", *Proceedings of the Royal Society B*, 276, 1161–1165.

Black, S., and Strahan, P.E. (2001). "The division of spoils: rent-sharing and discrimination in a regulated industry", *American Economic Review*, 91, 814–31.

Blau, F. and Kahn, L. (2016). "The Gender Wage Gap: Extent, Trends, and Explanations", IZA DP No. 9656.

Bloom, N., Propper, C., Seiler, S., and Van Reenen, J. (2015). "The impact of competition on management quality: Evidence from public hospitals", *Review of Economic Studies*, 82(2), 457-489.

Boggan, A., Bartlett, J., and Krawczyk, D. (2012). "Chess masters show a hallmark of face processing with chess", *Journal of Experimental Psychology: General*, 141(1), 37– 42.

Bronowski, J. (1973). *The ascent of man*, Little, Brown and Company, Boston.

Charness, N. (1992). "The impact of chess research on cognitive science", *Psychological Research*, 54(1), 4–9.

Charness, N., and Gerchak, Y. (1996). "Participation rates and maximal performance: A loglinear explanation for group differences, such as Russian and male dominance in chess", *Psychological Science*, 7(1), 46–51.

Croson R., and Gneezy, U. (2009). "Gender differences in preferences", *Journal of Economic Literature*, 47(2), 1–27.

Croizet, J., Després, G., Gauzins, M., Huguet, P., Leyens J., and Méot A. (2004). "Stereotype threat undermines intellectual performance by triggering a disruptive mental load", *Personality and Social Psychology Bulletin*, 30(6), 721-731.

de Bruin, A., Smits, N., Rikers, R., and Schmidt, H. (2008). "Deliberate practice predicts performance over time in adolescent chess players and drop-outs. A linear mixed models analysis", *British Journal of Psychology*, 99, 473–497.

de Sousa, J., and Hollard, G. (2015). "Gender differences: evidence from field tournaments", working paper, CEPREMAP.

Elo, A.E. (1978). *The rating of chessplayers, past and present*, Arco Pub, New York.

Ferreira, D.R. (2013). "The Impact of the Search Depth on Chess Playing Strength", *International Computer Games Association Journal*, 36(2), 67-80

Gerdes, C., and Gränsmark, P. (2010). "Strategic behavior across gender: A comparison of female and male expert chess players", *Labour Economics*, 17(5), 766-775.

Glickman, M. (1995). "A Comprehensive Guide to Chess Ratings", *American Chess Journal*, 3, 9–102.

Gneezy, U., Leonard, K.L., and List, J.A. (2009). "Gender differences in competition: Evidence from a matrilineal and a patriarchal society", *Econometrica*, 77(5), 1637-1664.

Gneezy, U., Niederle M., and Rustichini A. (2003). "Performance in competitive environments: gender differences", *Quarterly Journal of Economics*, 118, 1049–74.

Gneezy, U., and Rustichini A. (2004). "Gender and competition at a young age", *American Economic Review*, 94, 377-81.

Gränsmark, P. (2012). "Masters of our time: Impatience and self-control in high-level chess games", *Journal of Economic Behavior and Organization,* 82(1), 179-191.

Goldin, C., and Rouse, C. (2000). "Orchestrating impartiality: the impact of "blind" auditions on female musicians", *American Economic Review*, 40, 715–42.

Guid, M., and Bratko, I. (2006). "Computer analysis of chess champions", *International Computer Games Association Journal*, 29(2), 65-73.

Guid, M., and Bratko, I. (2011). "Using heuristic-search based engines for estimating human skill at chess", *International Computer Games Association Journal*, 34(2), 7181.

Günther, C., Ekinci, N.A., Schwieren C., and Strobel, M. (2010). "Women can't jump? An experiment on competitive attitudes and stereotype threat", *Journal of Economic Behavior and Organization*, 75, 395–401.

Gupta, N., Poulsen, A., and Villeval, M.C. (2013). "Gender matching and competitiveness. experimental evidence", *Economic Inquiry*, 51(1), 816–835.

Hambrick, D., Oswald, F., Altmann, E., Meinz, E., Gobet, F., and Campitelli, G. (2014). "Deliberate practice: Is that all it takes to become an expert?", *Intelligence*, 45, 34-45

Herlitz, A., and Lovén, J. (2013). "Sex differences and the own-gender bias in face recognition: A meta-analytic review", *Visual Cognition*, 21(9-10), 1306-1336.

Howard, R. (2004). "Are gender differences in high achievement disappearing? A test in one intellectual domain", *Journal of Biosocial Studies*, 37, 371–380.

Iriberri, N., and Rey-Biel, P. (2016). "Stereotypes are only a threat when beliefs are reinforced: On the sensitivity of gender differences in performance under competition to information provision", working paper, Barcelona GSE.

Jurajda, Š., and Münich, D. (2011). "Gender gap in performance under competitive pressure: Admissions to Czech universities", *American Economic Review Papers and Proceedings*, 101(3), 514-18.

Kleinjans, K. (2009). "Do gender differences in preferences for competition matter for occupational expectations?" *Journal of Economic Psychology,* 30, 701-710.

Krendl, A., Richeson, J., Kelley, W., and Heatherton, T. (2008). "The negative consequences of threat a functional magnetic resonance imaging investigation of the neural mechanisms underlying women's underperformance in math", *Psychological Science*, 19(2), 168-175.

Lavy, V. (2013). "Gender differences in market competitiveness in a real workplace: evidence from performance-based pay tournaments among teachers", *Economic Journal*, 123(569), 540–573.

Levitt, S., List, J., and Sadoff, E. (2011). "Checkmate: Exploring backward induction among chess players", *American Economic Review*, 101(2), 975-90.

Li, R. (2014). "Why women see differently from the way men see? A review of sex differences in cognition and sports", *Journal of Sport and Health Science*, 3, 155-162.

Maass, A., D'ettole, C., and Cadinu, M. (2008). "Checkmate? The role of gender stereotypes in the ultimate intellectual sport", *European Journal of Social Psychology,* 38, 231-245.

Morin, L-P. (2015). "Do men and women respond differently to competition? Evidence from a major education reform", *Journal of Labor Economics*, 33(2), 443-491.

Niederle, M., and Vesterlund L. (2007). "Do women shy away from competition? Do men compete too much?" *Quarterly Journal of Economics*, 122, 1067-101.

Niederle, M., and Vesterlund L. (2011). "Gender and competition", *Annual Review of Economics*, 3, 601-630.

Örs, E., Palomino, F., and Peyrache, E. (2013). "Performance gender-gap: Does competition matter?", *Journal of Labor Economics*, 31(3), 443-499.

Palacios-Huerta, I., and Volij, O. (2009). "Field centipedes*", American Economic Review*, 99(4), 1619-1635.

Paserman, D. (2010). "Gender differences in performance in competitive environments: evidence from professional tennis players", working paper, Boston University.

Polachek, S.W. (1981). "Occupational self-selection: a human capital approach to sex differences in occupational structure", *Review of Economics and Statistics*, 68, 60–69.

Robbins, T.W., Anderson, E.J., Barker, D.R., Bradley, A.C., Fearneyhough, C., Henson, R., Hudson, S.R., and Baddeley, A.D. (1996). "Working memory in chess", *Memory & Cognition*, 24, 83–93.

Rothgerber, H., and Wolsiefer, K. (2014). "A naturalistic study of stereotype threat in young female chess players", *Group Processes Intergroup Relations*, 17(1), 79-90.

Schwable, U., and Walker, P. (2001). "Zermelo and the early history of game theory", *Games and Economic Behavior,* 34(1), 123-137.

Shahade, J. (2005). *Chess bitch: Women in the ultimate intellectual sport*, Los Angeles, USA: Siles Press.

Shannon, C. (1950). "Programming a computer for playing chess", *Philosophical Magazine*, 41(314).

Shurchkov, O. (2012). "Under pressure: gender differences in output quality and quantity under competition and time constraints", *Journal of the European Economic Association*, 10(5), 1189–1213.

Steele, C. (1997). "A threat in the air: how stereotypes shape intellectual identity and performance", *American Psychologist*, 52(6), 613-629.

Vandegrift, D., and Brown, P. (2005). "Gender differences in the use of high-variance strategies in tournament competition", *Journal of Behavioral and Experimental Economics,* 34(6), 834-849.

Walton, G.M., and Cohen, G.L. (2003). "Stereotype lift", *Journal of Experimental Social Psychology*, 39, 456–467.

Wraga, M., Helt, M., Jacobs, E., and Sullivan, K. (2006). "Neural basis of stereotype-induced shifts in women's mental rotation performance", *Social and Cognitive Affective Neuroscience*, 2, 12–19.

Tables

Table 1: Descriptive statistics

	(1)	(2)	(3)	(4)
	Full sample		Switchers	
	Men	Women	Men	Women
p_{ig}	0.53	0.50	0.56	0.51
	(0.41)	(0.42)	(0.40)	(0.42)
Players's age	32.59	25.31	31.37	25.14
	(14.57)	(9.32)	(14.13)	(8.56)
Opponents's age	31.96	27.78	30.56	27.73
	(14.67)	(12.29)	(14.04)	(12.14)
Opponent is male	0.94	0.38	0.87	0.39
	(0.24)	(0.49)	(0.33)	(0.49)
Player's Elo	2370.06	2278.99	2409.62	2294.21
	(186.35)	(147.01)	(174.93)	(140.81)
Opponent's Elo	2333.34	2274.46	2347.98	2283.26
	(251.11)	(206.53)	(234.57)	(204.19)
Player's $\overline{error}$	16.48	17.71	15.63	17.30
	(20.74)	(21.73)	(19.88)	(21.27)
Opponent's $\overline{error}$	17.31	17.70	17.17	17.58
	(22.26)	(22.23)	(22.26)	(22.29)
Observations	50221	7715	23064	5695

Notes: The first column refers to the means and standard deviations for the full sample. The second column refers to the means and standard deviations for the sub-sample of players who play at least two games and play both men and women in our sample ('switchers').

Table 2: Gender and performance

	(1)	(2)	(3)	(4)
	Full sample	Switchers	Full sample	Switchers
Player is female	-0.010**	-0.010*	-0.023***	-0.021***
	(0.005)	(0.006)	(0.006)	(0.006)
Opponent is female			0.022***	0.021***
			(0.005)	(0.006)
Total effect				
Female-Female			0.001	0.001
			(0.005)	(0.006)
Games	57,936	28,759	57,936	28,759
R^2	0.212	0.217	0.213	0.218

Notes: The dependent variable is the number of points earned by i in the game: 1 for a win, 0.5 for a draw and 0 for a loss. The models are estimated by OLS on pooled data. Controls include the Elo differnetial, the ages of each player and the color i plays with. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 3: Is the gender of the opponent conditionally independent of the player's gender?

	(1)	(2)	(3)	(4)
	No controls	Switchers	+Female share	$+\overline{Elo_g}$
Player is male	0.563***	0.485***	0.020*	0.000
	(0.013)	(0.014)	(0.010)	(0.010)
Games	57,936	28,759	28,759	28,759
R^2	0.313	0.215	0.410	0.425
(Partial) ρ	0.559	0.463	0.247	0.000

Notes: The dependent variable is a dummy equal to 1 if the opponent is male. The share of non-i players who are female at the event is added in column (2). In column (3), we exclude players who only do not play against both men and women in our sample. In column (4), we use this same sub-sample and add the mean Elo rating of the player and opponent as a control. The models are estimated by OLS using pooled data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 4: Main results

	(1)	(2)	(3)	(4)	(5)
	Univariate	Controls	Switchers	Female players	Male players
Opponent is male	-0.102***	-0.040***	-0.040***	-0.043**	-0.042***
	(0.007)	(0.007)	(0.007)	(0.019)	(0.007)
Games	57,936	57,936	28,759	5,695	23,064
R^2	0.004	0.239	0.238	0.229	0.246

Notes: The dependent variable is the number of points earned by i in the game: 1 for a win, 0.5 for a draw and 0 for a loss. The models are estimated by OLS on within-i mean differenced data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 5: Robustness checks, different sub-samples I

	(1)	(2)	(3)	(4)	(5)
	No single sex events	No draws	≥20 games played	No Blitz or Junior	Event FE
Opponent is male	-0.041***	-0.044***	-0.032***	-0.041***	-0.039***
	(0.007)	(0.009)	(0.010)	(0.007)	(0.007)
Games	24,331	19,399	18,169	27,770	28,759
R^2	0.247	0.299	0.236	0.241	0.263

Notes: The dependent variable is the number of points earned by i in the game: 1 for a win, 0.5 for a draw and 0 for a loss. The models are estimated by OLS on within-i mean differenced data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 6: Robustness checks, different sub-samples II

	(1)	(2)	(3)	(4)	(5)
	|Elo dif|≤200	|Elo dif|≤100	|Elo dif|≤50	Elo≥2200	Elo≥2400
Opponent is male	-0.033***	-0.052***	-0.057*	-0.035***	-0.039***
	(0.010)	(0.017)	(0.030)	(0.008)	(0.010)
Games	19,534	9,587	4,585	23,795	14,407
R^2	0.150	0.079	0.069	0.242	0.235

Notes: The dependent variable is the number of points earned by i in the game: 1 for a win, 0.5 for a draw and 0 for a loss. The models are estimated by OLS on player within-i differenced data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 7: Non-linearity in the effect of Elo differential

	(1)	(2)	(3)	(4)	(5)
		$\overline{Elo_g}$ and $E[p_{ig}]$			
	Logged Elo ratings	Decile groups	Squares, cubes	Interacted	Intercept shift
Opponent is male	-0.045***	-0.038***	-0.036***	-0.030***	-0.034***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Games	28,759	28,759	28,759	28,759	28,759
R^2	0.228	0.239	0.242	0.247	0.239

Notes: The dependent variable is the number of points earned by i in the game: 1 for a win, 0.5 for a draw and 0 for a loss. The models are estimated by OLS on within-i mean differenced data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 8: Modeling the within-game quality of play

	(1)	(2)	(3)	(4)
	Women		Men	
	All	Switchers	All	Switchers
Opponent is male	0.087**	0.084*	-0.017	-0.001
	(0.041)	(0.043)	(0.020)	(0.020)
Games	7,715	5,695	50,221	23,064
R^2	0.022	0.024	0.005	0.008

Notes: The dependent variable is the logged mean error committed by i in between moves 15 and 30 of game g. The models are estimated by OLS on within-i mean differenced data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 9: Number of moves until resignation

	(1)	(2)	(3)	(4)
	Women		Men	
	All	Switchers	All	Switchers
j is male	-0.054	-0.033	-0.074***	-0.080***
	(0.037)	(0.040)	(0.021)	(0.022)
Games	2,229	1,605	12,801	5,274
R^2	0.131	0.150	0.060	0.067

Notes: The dependent variable is the number of moves of games ended with resignation. The models are estimated by OLS on within-i mean differenced data using only games which i lost. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 10: Quality of play and stakes

	(1)	(2)	(3)	(4)
	Women		Men	
Log stakes	0.034***		0.029***	
	(0.012)		(0.007)	
Opponent is male	0.079*	0.071	-0.000	-0.017
	(0.043)	(0.058)	(0.020)	(0.027)
Marginal effects				
Log stakes| opponent is female		0.030***		0.035
		(0.015)		(0.023)
Log stakes| opponent is male		0.042**		0.028***
		(0.019)		(0.007)
Games	5,695	5,695	23,064	23,064
R^2	0.011	0.026	0.030	0.030

Notes: The dependent variable is the logged mean error committed by i in between moves 15 and 30 of game g. The models are estimated by OLS on within-i mean differenced data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Appendix A: How chess engines determine 'best' moves

For the current paper, we use the Houdini 1.5a x64 program which according to the CCRL rating has an Elo rating of 3126, several hundred points above the very best human players in history. This rating is, however, based on games with other computer programs, and the question whether it is comparable to the ratings of human players is not easy to answer, due to the lack of matches where humans play against top performing computer matches under standard chess tournament conditions. There are nevertheless several indicators that top computer chess programs – including the one we use – are much stronger than any human player. In contrast to

humans, they: (1) have an enormous computing power, (2) use numerical values as evaluations, (3) adhere to the same rules all the time, and (4) are not influenced by emotions. Computer programs therefore have a capability of being more consistent than human observers, and can deal with incomparably more observations in a limited time. Since 2005, chess programs have defeated several grandmasters and even former world champions. In 2009, a computer chess program running on a mobile phone won a strong chess tournament "Copa Mercosur" with a performance rating of 2898 Elo points. The program searched about 20,000 positions per second, which is at least an order of magnitude less than the program that we use in this paper.

As mentioned in the main text, we consider 16 moves for each player between moves 15 and 30 of each game. Chess has an estimated state-space complexity of 10 to the power of 46. That is why we limit the search depth of Houdini to 15, meaning the program evaluates 15 moves forward from the move being evaluated. Such limits are necessary given the number of moves we wished to evaluate and the fact that going one move deeper often means double the computation required. Recent work in Ferreira (2013) estimates that Houdini 1.5a x64 at a search depth of 15 plies has an estimated Elo rating of 2563 (a grandmaster level), meaning there are players in our data who in fact have a higher Elo than the chess engine we use to evaluate moves. However, as noted by Guid and Bratko (2006), "even if [the computer's] evaluations are not always perfect, for our analysis they just need to be sufficiently accurate on average since small occasional errors cancel out through statistical averaging". Moreover, as noted earlier, the machines perform exceptionally well against human players. Nevertheless, we can show that our results are robust to the exclusion of players with Elo ratings in excess of 2563.

Houdini, as other chess programs, uses search trees to find and choose the best possible move. The root of the search tree is the player's current position on the board, nodes are chess positions, links between nodes represent chess moves, and leaves are the terminal positions of the tree. Because we limit the extent of computations, the leaves do not represent final positions, but the positions at the maximum depth of search.

When exploring search trees, Houdini determines the value of a starting point and chooses the best move. This exploration involves assigning evaluations to individual leaves and nodes in the constructed trees, employing a depth-first search. An evaluation function is used for this purpose, but only to assess the value of the leaves, while nodes are assessed on the basis of their immediate descendants, i.e. the leaves and nodes that their connections lead to.

Houdini employs the $mini - max$ concept, assuming that whatever is good for one player must consequently be bad for the other one. The root of a search tree is labeled as MAX, all nodes to which its connections lead as MIN, and then its descendants again as MAX, and so on. Evaluations are assigned as follows:

1. Nodes marked as MIN are given the lowest evaluation of their descendants.

2. Nodes marked as MAX are given the highest evaluation of their descendants.

The program thus evaluates, from the bottom up, all nodes in subtrees that are formed directly from the tree's root, which results in the final value of the root. This value is the ultimate assessment of the current position, and the move that leads to the root's descendant with the highest evaluation is chosen as the best move.

Appendix B: Supplementary quality of play analysis

Table 11: Supplementary quality of play analysis: Different sub-samples I

	(1)	(2)	(3)	(4)	(5)
	No single sex events	No draws	≥20 games played	No Blitz or Junior	Event FE
Female players					
Opponent is male	0.075*	0.095*	0.168***	0.079*	0.095*
	(0.045)	(0.050)	(0.058)	(0.043)	(0.050)
Games	4349	4022	3828	5542	5695
R^2	0.033	0.041	0.028	0.025	0.079
Male players					
Opponent is male	0.002	0.028	-0.042	-0.000	0.005
	(0.021)	(0.025)	(0.027)	(0.021)	(0.021)
Games	19982	15377	14341	22228	23064
R^2	0.009	0.016	0.013	0.009	0.043

Notes: The dependent variable is the logged mean error committed by i in between moves 15 and 30 of game g. The models are estimated by OLS on within-i mean differenced data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 12: Supplementary quality of play analysis: Different sub-samples II

	(1)	(2)	(3)	(4)	(5)	(6)
	|Elo dif|≤200	|Elo dif|≤100	|Elo dif|≤50	Elo≥2200	Elo≥2400	Elo<2536
Female players						
Opponent is male	0.084*	0.084*	0.084*	0.115**	0.181*	0.079*
	(0.043)	(0.043)	(0.043)	(0.050)	(0.093)	(0.044)
Games	5695	5695	5695	4158	1396	5475
R^2	0.024	0.024	0.024	0.025	0.046	0.024
Male players						
Opponent is male	-0.001	-0.001	-0.001	-0.025	-0.042	0.031
	(0.020)	(0.020)	(0.020)	(0.023)	(0.031)	(0.022)
Games	23064	23064	23064	19637	13011	16869
R^2	0.008	0.008	0.008	0.011	0.018	0.010

Notes: The dependent variable is the logged mean error committed by i in between moves 15 and 30 of game g. The models are estimated by OLS on within-i mean differenced data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.

Table 13: Supplementary quality of play analysis: Non-linearity in the effect of Elo differential

	(1)	(2)	(3)	(4)	(5)
		$\overline{Elo_g}$ and $E[p_{ig}]$			
	Logged Elo ratings	Decile groups	Squares, cubes	Interacted	Intercept shift
Female players					
Opponent is male	0.082*	0.077*	0.066	0.060	0.084*
	(0.043)	(0.043)	(0.043)	(0.043)	(0.043)
Games	5695	5695	5695	5695	5695
R^2	0.024	0.028	0.027	0.038	0.024
Male players					
Opponent is male	0.003	0.005	0.004	0.007	-0.001
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Games	23064	23064	23064	23064	23064
R^2	0.009	0.011	0.013	0.016	0.008

Notes: The dependent variable is the logged mean error committed by i in between moves 15 and 30 of game g. The models are estimated by OLS on within-i mean differenced data. Robust standard errors (in brackets) are clustered at the player level. Stars indicate statistical significance according to the following schedule: *** 1%, ** 5% and * 10%.