PERSONAL BESTS AND GENDER

Julio González-Díaz, Ignacio Palacios-Huerta, and José M. Abuín*

Abstract—We connect two large bodies of scientific inquiry. First, important theories in the social sciences establish that human preferences are reference-dependent. Second, a separate field of research documents substantial differences in preferences and attitudes across genders. Specifically, we examine the universe of official classic chess games (more than 250,000 subjects and 22 million games). This allows us to study differences across genders both in cognitive performance (intensive margin) and in competitive participation (extensive margin), using the fact that personal bests act as reference points. We find that males and females behave very differently around their personal bests in both margins.

I. Introduction

REFERENCE dependence captures the comparative nature of human feelings and perceptions. Important theories of economic behavior propose that human preferences are reference dependent. In models with these preferences, individuals evaluate outcomes relative to a reference point such as the current state (the status quo), past states, expectations about future states, or social comparisons. As such, reference-dependent preferences are at the heart of many behavioral models and concepts, including the endowment effect, loss aversion, status quo bias, and prospect theory. DellaVigna (2009), Bernheim, DellaVigna, and Laibson (2018), and O'Donoghue and Sprenger (2018) provide comprehensive reviews of theories, applications, and developments over the past two decades leading up to the current research frontier.[1] Interestingly, a conspicuous aspect of these reviews is that no part of what is a large body of research appears to be concerned with potential gender differences. This paper contributes to the existing literature by studying gender as a potential determinant of reference-dependent human behavior.

Understanding gender differences is in fact an area that has generated a voluminous, far-reaching literature in recent years, especially since Niederle and Vesterlund (2007). Croson and Gneezy (2009), Azmat and Petrongolo (2014), and Olivetti and Petrongolo (2016), for example, survey many studies on gender differences, including differences in risk preferences (e.g., emotions, attitudes, and overconfidence), in social preferences (e.g., by studying strategic situations in laboratory settings), and in competitiveness.

The results show that gender differences may have a huge impact on labor markets, and the family, and in human capital, consumption, investments, and many other decisions (Goldin, 2006, 2014). As such, these differences occupy a central place in the literature. And, yet, despite their prominent place, to the best of our knowledge no research has studied whether observed gender differences in behavior may stem from the *comparative* nature of human feelings and perceptions. Our goal is to study the link between these two influential bodies of scientific inquiry.

We take advantage of a setting that combines a number of unique characteristics, which we discuss in some detail in the next section. The setting (i) concerns a real-life cognitive task (in fact, the quintessential cognitive sport in humans: chess); (ii) is strictly competitive (zero-sum) with no potential elements of cooperation, and as such, it represents the cleanest possible context to study competitive behavior; (iii) involves a massive data set, specifically, the *universe* of officially rated classic chess games (more than 250,000 subjects and 22 million chess games) for two decades beginning from the year in which the world governing body the Fédération Internationale des Échecs (FIDE) first started publishing this information; (iv) contains detailed measures of performance and participation, which allows the study of both the intensive and extensive margins of effort; (v) concerns subjects who are experts, officially ranked, who have devoted a great deal of time to the task; and (vi) includes information on age, gender, and other demographic characteristics.

Besides these advantages, perhaps the main virtue is that, in this setting, reference points have been convincingly documented in the literature. In particular, Anderson and Green (2018) study subjects who play chess online and find that "personal bests" act as reference points. Small differences in outcomes are felt disproportionally around personal bests: players exert effort to set new personal best ratings and quit once they have done so. They further develop a loss-aversion effort model to substantiate this finding. These results are important and serve as the basis of our study, although, as we shall see, behavior around personal bests in official in-person competitions is different from that in online settings.

As anticipation of the results, we find that males and females behave very differently around their personal bests, in both the intensive and extensive margins of effort. In the extensive margin, women increase their effort more than men when approaching their personal best, but exert less effort after breaking it. In the intensive margin, women underperform relative to men—both to set a new personal best and after having done so.

As noted earlier, in terms of our contribution to the existing literature, no previous study appears to have linked

Received for publication November 20, 2020. Revision accepted for publication November 1, 2021. Editor: Shachar Kariv.

*González-Díaz: Universidad de Santiago de Compostela; Palacios-Huerta: London School of Economics and Ikerbasque UPV/EHU; Abuín: Universidad de Santiago de Compostela.

Financial support from the London School of Economics, FEDER, the Spanish Ministerio de Economía y Competitividad (ECO2015-66027-P, MTM2014-60191-JIN, and MTM2017-87197-C3-3-P), and the Dept. de Educación, Política Lingüística y Cultura del Gobierno Vasco (IT-869-13), and editorial assistance from Emily Faye Coles is gratefully acknowledged.

A supplemental appendix is available online at https://doi.org/10.1162/rest_a_01145.

[1] Recent substantial contributions to the literature include expectations-based reference points (Köszegi & Rabin, 2006, 2007, 2009) and models of salience (Bordalo, Gennaioli, and Shleifer, 2012, 2013, 2020).

the role of gender with the comparative nature of human behavior in a natural setting, in particular regarding cognition. As such, the gender differences in behavior around personal bests that we document may be taken as a contribution to our understanding of human nature.

The rest of the paper is structured as follows. Section II reviews other strands of related literature, including the advantages of the empirical setting. Section III goes over the data set, and section IV gives descriptive evidence. Section V presents the main results, section VI a brief discussion, and section VII concludes. Several robustness tests are provided in two online appendices.[2]

II. Related Literature

In addition to the literature on reference-dependent preferences and gender differences just discussed, this study also contributes to other strands of scientific inquiry.

A. Cognition and Human Capital

The fact that the setting involves cognitive performance is interesting per se, as understanding human cognition is essential in many areas. Numerous settings represent competitive situations that involve cognitive performance (e.g., test taking, student competition in schools, and competitions for promotion in firms and organizations). As such, discerning the determinants of cognitive performance is a relevant question in the literature on human capital, schooling, behavioral economics, and other areas. Also, research studies have established that measured cognitive ability (skills, effort, attitudes) is a strong predictor of occupational attainment, wages, and a range of social behaviors in adults, and several studies document its critical role in predicting the schooling performance of children, adolescents, and university students (see, e.g., Cunha et al., 2010; Heckman & Kautz, 2014). Besides social and economic outcomes, recent research also shows that cognitive ability is a main determinant of financial market outcomes. Also, intelligence (IQ) can be crucial in strategic situations, investments, and the formations of expectations; see, for example, Gill and Prowse (2016) and Proto, Rustichini, and Sofianos (2019) and other references therein. This paper contributes to these fields by studying reference points (personal bests) as a potential determinant of cognitive performance and participation in competitive environments.[3]

B. Endogenous Preferences

Rayo and Becker (2007) argue that evolution favors a happiness function that measures the individual's success in *relative* terms, and "an individual's utility, whether defined in terms of decision making or hedonic experience, tends to be sharply influenced by his personal history" (pp. 302–303), which intuitively includes his or her own personal best in a prominent role. In their analysis, happiness functions are based on a measure of success relative to a "performance benchmark" or reference point, and the difference between a person's output and that point is the carrier of happiness. We adopt the same viewpoint here. Also, an important literature studies the extent to which the nature and formation of preferences is susceptible to direct influences from the social and economic environment, as well as the consequences that this endogenous relationship may have for behavior.[4] Our study relates to and extends existing research by examining behavior around a reference point that is not based on rational expectations and is endogenous to the economic agent. Further, it relates to the literature on intrinsic and extrinsic incentives on which there is significant theoretical work in economics, as we study incentives, gender, and personal bests (see, e.g., Benabou & Tirole, 2002, 2003, 2004).

C. Goals

Research in psychology suggests that "goals" may serve as reference points (see, e.g., Allen et al., 2017; Heath, Larrick, and Wu, 1999; Williams & Gilovich, 2012; and Pope & Simonsohn, 2011). In economics, theoretical and experimental research is also beginning to study goals more closely (see, e.g., Smithers, 2015 and Koch & Nafziger, 2016).[5] Importantly, Damon et al. (2020) finds a relevant role for goals by manipulating them in field experiments with students where they ask subjects to set goals for themselves, both task-based and performance-based. Personal bests, however, are a natural reference point, whether or not they act as goals (Anderson & Green, 2018). Further, in our natural setting we see no intervention by experimenters, information about one's own personal best is readily available and salient for each subject, and we can document the intensive and extensive margins of the mechanism at play.

D. Sports as a Field Economics Lab

In recent years, sports settings have proven to be a useful data-rich, lablike setting that is able to inform economics in insightful ways. The reason is that important elements of human behavior are sometimes starkly observable in these settings. Often good data are abundant, the goals of the participants are precisely determined, the outcomes are extremely clear, the stakes are high, and the subjects are professionals

[2] Appendix A is available online at https://bit.ly/PBGenderAppA, and appendix B at https://bit.ly/PBGenderAppB.

[3] In the literature on cognitive performance, González-Díaz and Palacios-Huerta (2016) study the impact of competition dynamics using a natural experiment in chess competitions, and in the literature on gender differences Dilmaghani (2020) studies the role of time constraints in performance in chess tournaments.

[4] See, for example, Becker (1996), Becker and Murphy (2000), Palacios-Huerta and Santos (2004), Bowles (1998), and Fehr and Hoff (2011) and other references therein. See also Dawson and de Meza (2018) for an evolutionary explanation of different behavioral biases.

[5] Although not focused on gender differences, and obviously not cognitive, see Markle et al. (2018) and Burdina, Hiller, and Metz (2017) on marathon runners and Harding and Hsiaw (2014) on goals for energy conservation.

with experience. Thus, not surprisingly, a number of prominent findings in economics have been documented for the first time in sports settings.[6]

With respect to the specific advantages of our natural setting, in this paper we use data from the most popular cognitive sport in the history of humankind: chess. Much like other sports settings, it represents a valuable opportunity for studying an open question in the literature because of a number of useful characteristics.

First, chess is a complete information game that involves no chance elements. The game is zero-sum (or strictly competitive), with no potential elements of cooperation. As such, it is the cleanest possible context to study competitive behavior.

Second, the study concerns high-stakes decisions that subjects are familiar with, that really affect them, to which they are accustomed, and that take place in their own real-life competitive environment. From the perspective of observing and measuring behavior, a comprehensive data set is available where choices, outcomes, and other characteristics are cleanly measured.

Third, the setting concerns subjects with a high level of expertise and skill levels. It ranges all the way up to the highest possible degree of human cognitive skills at the task under study, including the best players in the history of the game, such as Magnus Carlsen and Garry Kasparov.

Finally, a usual difficulty in the study of cognition is that measures of cognitive abilities are often lacking in the literature or at best can be measured indirectly. Here, however, we can find a highly precise measure of the cognitive ability of the players at the task they perform. As discussed below, subjects are rated according to the *ELO rating* method, a measure that provides close estimates for the probability that one player will outperform the other at the cognitive task.

We refer throughout the paper to "effort," which we take as the amount of costly resources invested into the effective development of cognitive skills. In terms of consequences, we consider performance and participation as the intensive and extensive margins of this effort. The inputs that go into performance may include, for example, time studying chess, preparing, practicing with others, and all types of costly activities that improve cognitive performance. As for the decision to play, inputs include those associated with competitive participation given that, for skills to be effectively developed, subjects have to compete in official tournaments. Thus, inputs include traveling costs, tournament fees, and opportunity cost of time.

[6]For instance, without attempting to be exhaustive, Ehrenberg and Bognanno (1990) investigate tournament incentive effects in golf tournaments, Szymanski (2000) studies discrimination using soccer data, Palacios-Huerta (2003, 2014) tests the implications of the Minimax theorem, Garicano, Palacios-Huerta, and Prendergast (2005) study social pressure as a determinant of corruption in professional soccer, Romer (2006) analyzes optimal decision-making using football data, González-Díaz, Gossner, and Rogers (2012) look at heterogeneity in high-stakes performance in tennis, Palacios-Huerta and Volij (2009) address backward induction using chess players, and Pope and Schweitzer (2011) study loss aversion using evidence from golf.

III. Empirical Setting and Data Set

The data set comes from the FIDE, the world governing body of chess. It contains information about the *universe* of the official classic chess games that are played in official tournaments valid for the computation of the ELO rating (the official measure used to rank chess players and evaluate their performances), which the FIDE started publishing publicly online in 2000. Essentially, the difference between the ELO ratings of two players in a chess game is functionally related to an estimate of the probabilities that they beat each other. After a game, the winner gets some rating points from the loser, a number of points that depends on their rating difference. In case of a draw, the lower-rated player also gains certain points from the higher-rated one. The *performance* of a player in a game or series of games (say, in a given chess tournament) is computed using a formula that depends monotonically on the average rating of the opponents he or she has faced and the scores he or she has obtained against them.[7]

Nowadays, the top ten players in the world typically have an ELO rating between 2,770 and 2,865 points, with the top hundred players a rating above 2,650 points. Players with a rating above 2,500 points are in general professionals who have the title of Grandmaster, which is the highest title that a player can achieve. Strong club players have above 2,000 points. Magnus Carlsen, ranked number one in the world, currently has 2,855 points (November 2021).

Every month FIDE publishes the *FIDE rating lists*, which contain the updated information about the ELO ratings of more than 250,000 players worldwide and, more importantly, about the games they have played in the last month (from the previous list to the current list). These lists include their results and the ratings of the opponents they have faced, plus a set of demographic characteristics for each player.[8] Needless to say, performance in a given game or set of games can be above or below the rating. For example, the 82nd Tata Steels Masters tournament in Wijk Aan Zee (the Netherlands, January 10–26, 2020) was won by Fabio Caruana, currently ranked number two in the world. His rating at the beginning of, and during, the tournament was 2,822 ELO points (January 2020 list), but his performance *during* the tournament was 2,945 ELO points (he played during the tournament *as if* he had that rating). As a result, his updated ELO rating in the February 2020 list was 2,842 ELO points. He played with this rating until he played more games, and his rating was updated in a new list.

[7]For a detailed explanation of the ELO Rating method we refer to Chapter B.02 in the FIDE Handbook (2017, https://handbook.fide.com/).

[8]From January 2000 to July 2009, rating lists were published at quarterly frequencies, from July 2009 until July 2012 every two months, and since July 2012 every month.

As noted earlier, a main basis for our study is Anderson and Green (2018), who study subjects playing chess online. This is a nonofficial, unregulated, low-stakes setting that typically involves all types of chess aficionados. It has no information on gender, age, official FIDE ratings, or demographic characteristics. Consistent with their findings, we first confirm and take as a reference point the maximum rating *ever* of a player (his or her "personal best") and then study how behavior depends on the distance to it.

The complete FIDE data set has a large variety of players, including many who have participated in very few official tournaments during the past two decades.[9] In our analysis we require a minimum experience in official tournaments of at least 50 official chess games.[10] This minimum results in a data set with a total of 103,761 players who have played 14,028,274 games and have been observed in 6,854,205 rating lists. The average rating in the data set is 2,001 ELO points, which corresponds to a strong club player. Roughly 90% of the rating lists correspond to men, who also happen to represent 90% of the players. The average rating is 2,013.6 ELO points (standard deviation [std. dev.] 244.5) for men and 1,882 ELO points (std. dev. 266.9) for women. The average age is 41.29 (std. dev. 17.77) for men and 26.98 (std. dev. 13.53) for women.

Players do not typically play in official tournaments continuously over the year. When a player is "active" in a given period, that is, when he or she plays some games since the previous list was published, we will say that he or she has an *active list*. The rest of his lists are called *inactive lists*. The percentages of active lists are 27.1% for men and 26.4% for women. Among men, young players under 20 years of age represent 12.8%, players aged 20 to 60 represent 71.1%, and players over 60 represent 16.1%. For women, these percentages are 33.8%, 62.1%, and 4.1%, respectively.

Finally, the personal best has been broken in 269,772 rating lists, that is, in roughly 3.9% of all the lists (or 14.5% of the active lists). Overall, 54.6% of these breaks correspond to players under 20 years of age, 43.5% to players between 20 and 60, and 1.9% to players over 60. Across genders these percentages are 51.3%, 46.6%, and 2.1% for men, and 78.3%, 21.4%, and 0.3% for women.

IV. Descriptive Evidence

In this section we begin with a visual description of subjects' activity (in terms of games and types of rating lists), as well as their performance by gender, as a function of the distance to their personal best. We then focus on performance split by age groups.

Figure 1 provides some initial insights into differences in behavior across genders. Empty circles correspond to men and filled circles to women. The first row represents aggregate information on various characteristics of the rating lists (number and types of lists, percentage of active lists, monthly games per list, and performance since the previous list was published). Omitting players' identities, we denote by r_t the rating of a player in the list published at time t, by $n_{t,t+1}$ the number of games he played from t to $t+1$ with that rating, and by $p_{t,t+1}$ the performance in those games. We use $pbest_t = \max_j\{r_j,\ j \leq t\}$ to denote the personal best rating ever reached by that player up to time t.

Panel 1A shows the total number of rating lists corresponding to each gender as a function of the distance to the maximum rating of the player, $dist_t = r_t - pbest_t$. Obviously, $dist_t \leq 0$. It also includes how many are active lists and inactive lists. We observe clear differences across genders, driven, of course, by the gender proportions in each type of list. We also see that for men there seems to be a peak of observations around 20 rating points away from their personal best, but not for women. Panel 1B shows the average number of monthly games played. Getting closer to the personal best appears to correlate with greater activity, which is intuitive. Interestingly, men and women have similar activity levels when far from the maximum rating, whereas women are more active than men when getting closer to their personal best. Panel 1C provides another measure of players' activity: the percentage of lists in which a player is active. The pattern has some similarities with that in panel 1B: women, who now appear to be less active than men when far away from their personal best, increase substantially their activity level when approaching it. Panel 1D reports average performances.[11] Average performance tends to increase as the personal best is approached, which again is intuitive. Also, consistent with the raw ELO ratings data noted earlier, the average rating of women is lower than the average rating of men at any distance from the maximum rating. More importantly, there seems to be a decline in the average performance of women as they approach their personal best (within a 20–30 point distance), whereas we see no apparent decline for men.[12]

The second row of panels provides some initial insights into dynamic behavior by studying two consecutive rating lists. This generates the possibility that from one rating list to the next one some subjects may have beaten their personal best. The idea is to study *intertemporal* changes in activity and performance, depending on whether or not the personal

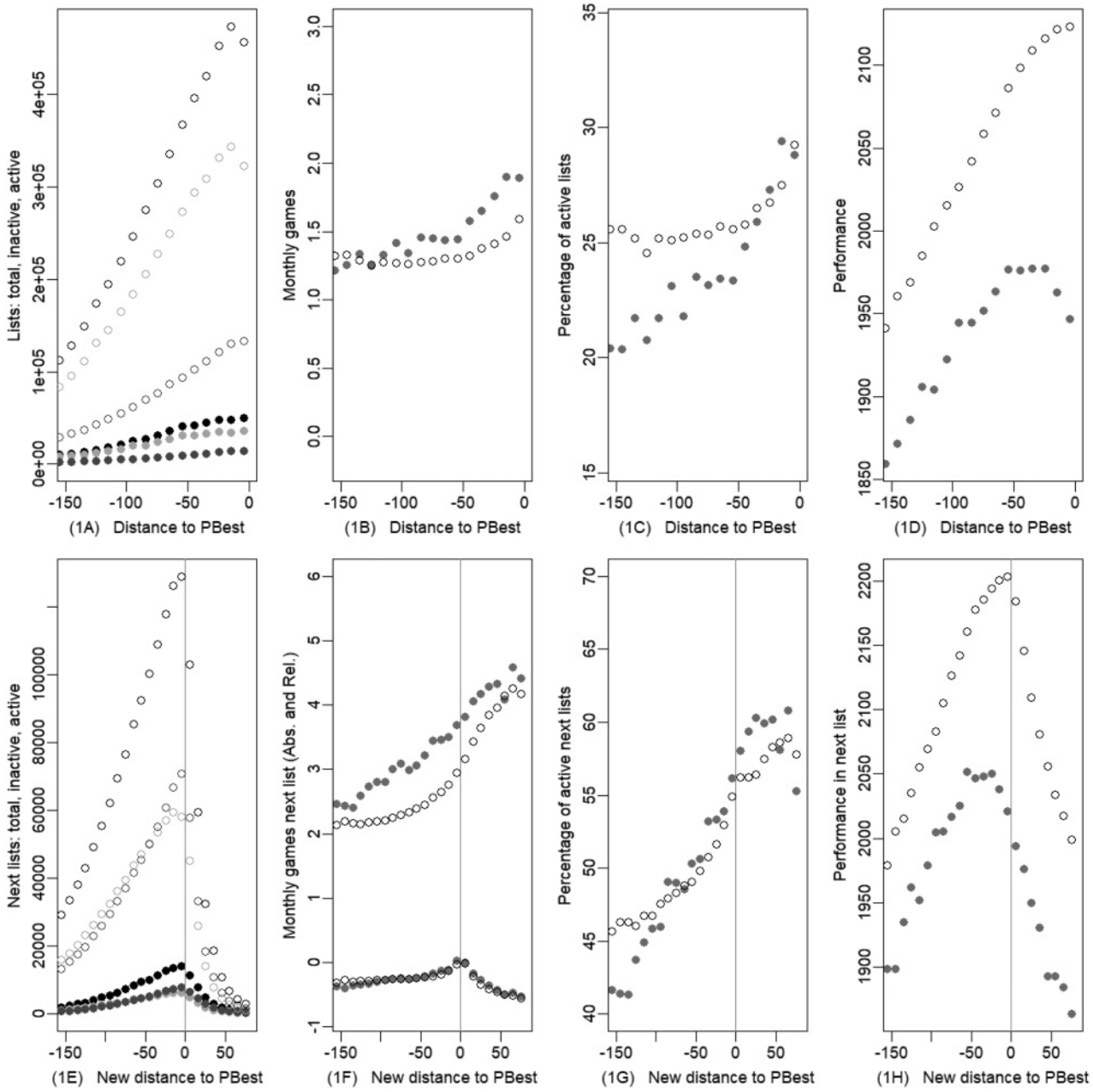
[9]For instance, approximately 42,000 players have played no officially rated game since 2000, and many are young players who stop playing at around the age of twenty. Our data set spans from January 2000 to March 2018.

[10]This is roughly equivalent to having played a minimum of five official tournaments (about 50 days in total) in two decades. The results are robust to changes in this minimum. In the online appendices we include a computer script that allows the implementation of the same analyses we report by varying the minimum level of experience.

[11]To measure performance in a meaningful way we consider rating lists in which a player has a minimum number of five games. The results are robust to changes in this minimum number.

[12]Anderson and Green (2018) find that among online chess aficionados, average performance minus average online rating tends to decrease as the personal is being approached until they are within ten rating points from the personal best, when it tends to increase.

FIGURE 1.—ACTIVITY (GAMES AND LISTS) AND PERFORMANCE BY GENDER WITH DISTANCE TO PERSONAL BEST



• Women ○ Men. Women's lists: ● ● ● Total, inactive, active. Men's lists: ○ ○ ○ Total, inactive, active.

best is broken. Consider players who are active from t to $t+1$ (green lines in panel 1A). Performance in the games played during this period could in principle be such that the new rating at $t+1$ is above the previous personal best rating. The x-axis reports the distance from the *new* rating at $t+1$, r_{t+1}, to the personal best standing at t. Observations to the right of zero represent the lists in which the maximum is broken in that period.

Panel 1E shows the number of total, active, and inactive lists from $t+1$ to $t+2$ (which we call total, active, and inactive rating lists at $t+1$), conditional on $n_{t,t+1} > 0$. The number of observations to the right of zero is small and decreasing, which indicates that we find few rating lists in which a player has just beaten his or her personal best, especially by a large amount.

Panel 1F reports two interesting aspects. The upper lines show the monthly games played $n_{t+1,t+2}$, conditional on $n_{t,t+1} > 0$.[13] These lines show a positive correlation between

[13]Note that these lines start slightly above 2, whereas in panel 1B they start slightly above 1. The reason is that these are rating lists conditional on a player being active, whereas in panel 1B they are unconditional. As players tend to be active or inactive during the year for periods that span several rating lists, it is natural that, conditional on being active, more activity occurs in the next list.

activity and setting a new personal best, both before and after. Needless to say, this correlation does not mean or imply anything about causality. To get additional insights, we also include the *growth rate in activity* which is defined as $(n_{t+1,t+2} - n_{t,t+1})/n_{t,t+1}$. These are the two additional series at the bottom of the panel. We see that this growth rate increases as the players get closer to the maximum rating (as above), but that after breaking the personal best, and contrary to the top lines, it decreases. This suggests a decrease in activity after setting a new personal best.[14] Panel 1G reports the percentage of active lists, showing similar patterns to the upper lines in the previous panel, now with a somewhat greater level of activity by women after breaking their personal best. Finally, panel 1H reports performances $p_{t+1,t+2}$, conditional on $n_{t,t+1} \geq 5$.[15] We find again a decline in women's performance as they approach their personal best (but not in men's performance). We also observe a sharp decline in average performances, similar for both men and women, after they set a new personal best. This is intuitive since the lower the rating, the easier it is to break the maximum. As this panel is unconditional on any other characteristics, it again simply suggests potentially interesting differences across genders.

Figure 2 focuses on players' performances in the complete sample and for two large age groups: young players (under 20) and adult players between 20 and 60 years of age.[16] The first two columns of panels concern ratings and performances, and the last two columns concern two consecutive rating lists as in the bottom panels of the previous figure.

Panel 2A reports performances. It corresponds to panel 1D in the previous figure, but we have also added the average rating of the players (solid lines). We find that for women (but not for men), their ratings tend to be below their performances.[17] Panels 2E and 2I split the sample by age groups. Young players, regardless of their gender, tend to perform above their ratings, which is intuitive because younger players are typically those more rapidly improving. On the other hand, subjects aged 20 to 60, both men and women, tend to perform slightly below their ratings, especially women when they are close to their personal best.

Panel 2B reports performances relative to own ratings, denoted as *relative performance* and computed as $(p_{t,t+1} - r_t)/r_t$. Consistent with the previous discussion, women tend to overperform relative to their rating in a somewhat stable fashion as a function of the distance to their personal best. Men tend to underperform, but increasingly less so as they approach their personal best. Indeed, both genders appear to overperform their ratings quite similarly when they are close to their personal best, whereas they perform quite differently when away from it. Across age groups we also observe interesting differences (panels 2F and 2J): a clearly decreasing overperformance that is similar for both males and females when young, whereas for subjects aged 20–60 we see a pattern of underperformance in men (again, increasingly less so as they approach their personal best), and a much less clear pattern in women.

Similar to the bottom panels in the previous figure, panels 2C and 2D deal with dynamic performance in consecutive periods, which may include cases in which a new personal best rating has been set. Panel 2C reports $p_{t+1,t+2}$ conditional on $n_{t,t+1} > 0$, as in panel 1H in the previous figure, and it also includes players' ratings r_{t+1}. There seems to be a slight tendency in men to have a rating above their performance, and in women to have a performance above their rating. Panels 2G and 2K show intuitive relationships between performance and ratings across age groups. Performance is above ratings for young players, who are more rapidly improving, and similar to ratings in the older age group. Setting a new personal best changes the patterns of both performances and ratings. Finally, panels 2D, 2H, and 2L report relative performances defined as $(p_{t+1,t+2} - r_{t+1})/r_{t+1}$. The patterns after breaking the personal best (to the right of zero) seem to continue those before breaking the personal best. This is clearly the case in panel 2H for the younger age group with a decreasing pattern of overperformance that continues after setting a new personal best. For the main age group 20–60 (panel 2L), relative performances are clearly negative and quite stable for men, while for women they again exhibit a decreasing tendency.

Summing up, the raw data suggest a number of interesting and potentially important gender differences in both the intensive margin (performance) and the extensive margin (participation and intertemporal activity) around personal bests. Needless to say, it is not possible to draw deep conclusions before implementing a more rigorous analysis. We turn to this analysis next.

V. Main Results

We divide this section into two subsections. In the first one, we study subjects' behavior as a function of the distance to their personal best. Subjects are always at a point where $dist_t \leq 0$. In the second one, we are interested in their

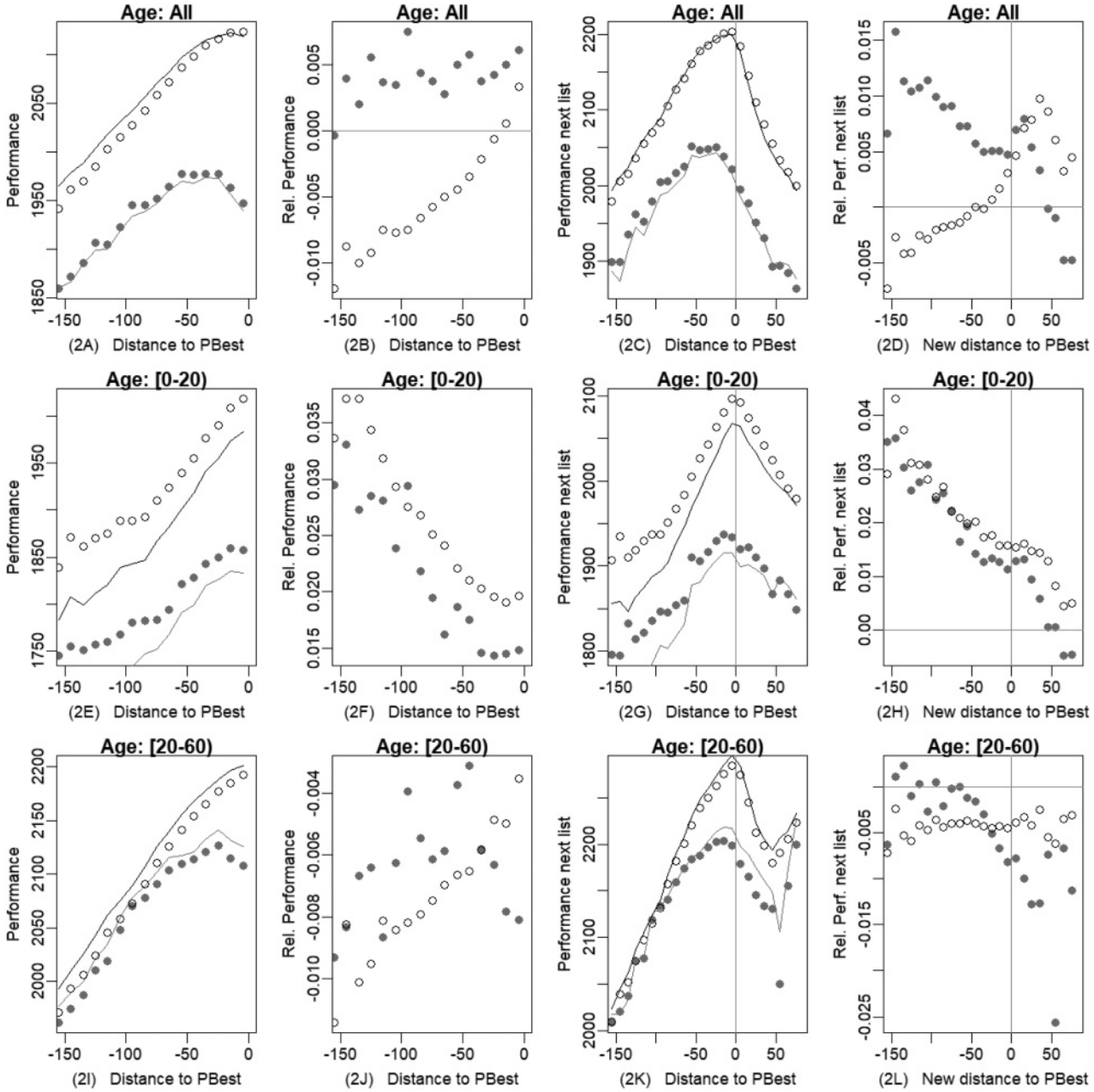
[14] Anderson and Green (2018) find a stable probability of quitting playing online chess (not playing for at least 1 hour) before setting a new personal best, and a discontinuous jump up in this probability after setting a new personal best.

[15] We choose $n_{t,t+1} \geq 5$ to make it readily comparable with the top panel 1D because, as indicated earlier, this is the minimum number of games used to compute *performance*.

[16] Different cut points around twenty and sixty do not change the patterns that are observed.

[17] This is likely driven, at least in part, by the larger proportion of games by young players among women. The design of the original formula to compute rating variations was such that the distribution of the expected rating variation was, for all players, symmetrically centered around zero. Thus, the Law of Large Numbers implies that for a large sample of players with constant playing strengths, the rating and performance lines should overlap almost perfectly. Yet some adjustments were later made in the formulas implemented by the FIDE, which combined with the impact of young, improving players may explain part of the difference between the two lines. In particular, the so called "400-point rule" states that any rating difference between two players beyond 400 points (which is a very large and somewhat unusual by chess standards) is taken to be precisely 400 points. See the FIDE Handbook (2017) for details.

FIGURE 2.—PERFORMANCE AND RATING BY AGE AND GENDER



• Women ○ Men.

intertemporal behavior, in particular in how they behave *after* they set a new personal best.

A. *Distance to Personal Best: Effect on Performance and Activity*

As in the descriptive section, performance is computed only in subjects' lists with a minimum activity level (five games). We also drop a small number of observations (0.24%) in which performance cannot be computed because the average rating of the opponents is not available. Table 1 reports two sets of regressions. In panel A, the dependent variable is relative performance $(p_{t,t+1} - r_t)/r_t$.[18] Part I in this panel reports OLS regression estimates. In column 1, we include only the gender variable *Female*. In the second specification, we add a dummy variable indicating whether or not the personal best is within reach of a 10-point distance (*Reach10*) and the *Age* of the player. In the third we add an interaction term between *Female* and *Reach10* but drop *Age*. The last two specifications are the most complete ones. In particular, in the last one we include in addition to

[18] We have implemented the same regressions in this and the next subsection using absolute performance instead of relative performance, finding essentially no qualitative differences. See appendix B.

TABLE 1.—DISTANCE TO PERSONAL BEST: PERFORMANCE

Panel A—Dependent variable: p_t/r_t					
Part I: OLS Specifications	(1)	(2)	(3)	(4)	(5)
Constant	−2.599***	−25.270***	−5.708***	25.090***	86.480***
	(0.063)	(0.167)	(0.073)	(0.169)	(0.611)
Female	6.907***	−3.267***	9.480***	−2.304***	−3.487***
	(0.195)	(0.197)	(0.235)	(0.238)	(0.240)
Reach10		0.743***	11.750***	−1.221***	−1.349***
		(0.135)	(0.142)	(0.154)	(0.157)
Female × Reach10			−10.080***	−2.956***	−4.555***
			(0.418)	(0.413)	(0.411)
Age		−0.726***		−0.724***	−0.715***
		(0.003)		(0.003)	(0.003)
PBest					−0.031***
					(0.000)
PBestDuration					0.015***
					(0.000)
Titles fixed effects	No	No	No	No	Yes
Adjusted R^2	0.000	0.042	0.008	0.042	0.048
N (lists)	1,211,051	1,211,051	1,211,051	1,211,051	1,211,051
N (games)	12,535,967	12,535,967	12,535,967	12,535,967	12,535,967
Part II: OLS coefficient (std. dev.) of different Female × Reach interactions					
Female × Reach10			−10.080***	−2.956***	−4.555***
			(0.418)	(0.413)	(0.411)
Female × Reach20			−9.524***	−2.454***	−4.174***
			(0.397)	(0.392)	(0.391)
Female × Reach50			−8.738***	−1.896***	−3.603***
			(0.399)	(0.394)	(0.392)
Female × Reach100			−7.158***	−1.372***	−2.890***
			(0.510)	(0.502)	(0.500)
Panel B—Dependent variable: New personal best is set Coefficient (std. dev.) of different Female × Reach interactions in logit regressions					
Female × Reach10			−0.683***	−0.329***	−0.298***
			(0.016)	(0.017)	(0.017)
Female × Reach20			−0.793***	−0.431***	−0.390***
			(0.019)	(0.019)	(0.020)
Female × Reach50			−1.096***	−0.630***	−0.529***
			(0.033)	(0.034)	(0.034)
Female × Reach100			−1.412***	−0.760***	−0.592***
			(0.088)	(0.088)	(0.089)

In panel A, relative performance is multiplied by 1,000. Specifications in Part II correspond to columns 3, 4, and 5 in Part I, just varying Reach to consider different distances from the personal best. Only coefficient estimates (std. dev.) for the interaction term are reported. Similarly, in Part B where the dependent variable is equal to 1 when a new personal best is set. *, **, *** denote significance at 5%, 1%, and 0.1%.

all the previous conditioning variables: the maximum rating the player has achieved (*PBest*), how long the personal best has been standing (*PBestDuration*), and fixed effects for the official titles (*Titles*) the player has achieved.[19]

Consider the two most complete specifications. Consistent with intuition, age is relevant and has a stable impact across specifications. The older the player is, the harder it is to overperform relative to the player's own rating. Also as expected, the higher the personal best is, the harder it is to perform above one's own rating. Players with high best ratings and older players are intuitively more "established," whereas it is easier to make substantial improvements when young and when the rating is lower. *Reach10* is also significant with a negative coefficient. This is intuitive. When players are closer to their maximum ratings (rather than when they are farther away from it), they should intuitively find it harder to overperform relative to their own rating.

With respect to gender differences, the specifications that do not take into account the age of the player would appear to suggest that, relative to men, women overperform more easily their own ratings. These specifications are of limited interest though, because large differences are seen in age distributions across gender pools. Once *Age* is included in the regression framework, the results indicate that women underperform relative to men. This underperformance is strongly significant at conventional significance levels and quite stable in magnitude. Importantly, the interaction term between *Female* and *Reach10* is also negative and strongly significant. The magnitude of this coefficient is also quite

[19]Essentially three official titles are awarded by the FIDE: Grandmaster (GM), International Master (IM), and Federation Master (FM). The GM title is the highest title a chess player can achieve. The title IM ranks below the GM title, and the FM title below the IM title. The requirements for achieving one of these titles are somewhat complex. They involve achieving a prespecified ELO rating and obtaining certain outcomes in certain tournaments. Similar titles also exist that apply only to women. Current regulations may be found in the official FIDE Handbook (FIDE, 2017).

large relative to all other variables. Consistent with the intuition in the raw data, this interaction indicates that women underperform relative to men, especially when they are close to the personal best (within ten rating points).[20] We next study this effect more closely.

The second part of panel A reports the coefficient estimates just for the interaction term between *Female* and *ReachX* in the last three regression specifications when considering three alternative values of the distance to the personal best: *X* equal to 20, 50, and 100 rating points, respectively. We find that the effect is always negative and strongly significant. Importantly, the magnitude decreases in absolute terms when subjects are farther away from the personal best, as could be expected. For example, for a player with an ELO rating of 2,000, the effect goes from a decrease in performance of $2 \times 4.55 = 9.1$ points (*Reach10*) to a decrease in $2 \times 2.89 = 5.7$ points (*Reach100*).

Panel B studies the determinants of performance using a different regression framework. The dependent variable now takes the value 1 when the personal best is broken ($r_{t+1} > pbest_t$) and 0 otherwise. We then implement logit regressions for the same three specifications of the independent variables as in panel A and report the coefficient estimates for the interaction term between *Female* and *ReachX* for the same four values of the distance to the personal best. The results are consistent with those in panel A. In particular, this interaction effect is negative and continues to be strongly significant. Interestingly, the magnitude of the effect (absolute value) increases with the distance to the personal best in each of the specifications. This suggests that women, relative to men, become increasingly less likely to beat their personal best as the distance to it becomes larger. In terms of magnitudes, the coefficients translate from a probability of about 0.55 (*Reach10*) to about 0.38 (*Reach100*).[21] In appendix A we explore in more detail this effect and provide additional evidence. We also report the results of several additional specifications, including different minimum experience levels and separate regressions for different age groups. Some minor differences may be of interest in their own right, but the basic results are quite robust. In particular, as could be expected, we find essentially identical results for the more numerous and main age group of interest of subjects aged 20 to 60.[22]

[20]Title fixed effects are often significant but do not impact much the magnitude and significance of all other variables.

[21]In the literature men are typically found to be more competitive *versus others* than women. Interestingly, these results show that when competing *against oneself,* women may be more competitive than men. See Croson and Gneezy (2009) and Palacios-Huerta (2021) and other references therein.

[22]For the young group of subjects under 20 years of age (AgeU20), we find in the panel A regressions that the interaction effect remains strongly significant, though the magnitudes are smaller. Interestingly, for the logit regressions in panel B the interaction becomes essentially zero, and not significant at conventional significance levels. For the older group of subjects above 60 years of age (Age60+), the interaction term changes sign in both panels and has little significance.

We next study the extensive margin: activity. The dependent variable in table 2 is the number of monthly games that subjects choose to play.

We study the same specifications as in the previous table. In the last two most complete specifications, we find that the younger a player is, and the lower his or her rating is, the greater is the activity level.[23] *Reach10* is positive and significant at conventional levels. With respect to gender differences, we find two noteworthy effects. First, women tend to be less active than men, but, interestingly enough, when the personal best is "within reach" they are significantly more active than men. This effect is quite large in magnitude. Second, in the bottom part, we find that this interaction is always positive and strongly significant and shows a tendency to increase with the distance to the personal best when moving from within 10 points, to within 20, 50, and 100 rating points. Specifically, with a baseline of 3.33 games per month, this effect goes from 0.40 more games (*Reach10*) to 0.47 more games per month (*Reach100*). We note here that 100 rating points is a substantial distance in chess. In appendix A we again find that these main results maintain for the main group of subjects aged 20 to 60. As for the younger subjects under 20 years of age, the interaction effect operates in the same direction, remains quite significant, and is larger in size: young women are more active than men overall, and especially so when the personal best is within reach.

Summing up, the evidence is consistent with differential responses by gender, in terms of both performance and activity levels. In particular, the results are consistent with the hypothesis that the distance to the personal best is associated with a greater increase in activity and with a decrease in performance in females relative to males.

B. Setting a New Personal Best: Effect on Performance and Activity

We next study gender differences in intertemporal behavior. We study behavior from $t+1$ to $t+2$ conditional on having been active the previous period, $n_{t,t+1} > 0$. This opens up the possibility that some subjects may have set a new personal best from t to $t+1$. We first study relative performance and then changes in playing activity. This conditional behavior means that the number of (pairs of) lists decreases to about 420,000 and 812,000 lists, respectively. In terms of chess games, we now study 5.1 and 6.6 million games. We introduce a new variable, *PBestBroken,* which takes the value one when the personal best is broken from t to $t+1$, that is, when $r_{t+1} > pbest_t$.

The dependent variable in table 3 is $(p_{t+1,t+2} - r_{t+1})/r_{t+1}$. As in table 1, we find that the older a player is, and the higher his or her personal best is, the harder it is to overperform relative to the player's own rating. Likewise, being within a 10-point reach of the personal best and its

[23]Note that the last specification includes Titles fixed effects, whose impact can be seen in appendix A.

TABLE 2.—DISTANCE TO PERSONAL BEST: ACTIVITY

Dependent variable: Monthly games played					
Part I: OLS specifications	(1)	(2)	(3)	(4)	(5)
Constant	1.147***	2.057***	1.360***	2.074***	3.339***
	(0.001)	(0.003)	(0.001)	(0.003)	(0.012)
Female	0.234***	− 0.003	0.141***	− 0.108***	− 0.056***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)
Reach10		0.352***	0.562***	0.303***	0.573***
		(0.003)	(0.003)	(0.003)	(0.003)
Female × Reach10			0.326***	0.450***	0.404***
			(0.010)	(0.010)	(0.010)
Age		− 0.015***		− 0.016***	− 0.020***
		(0.000)		(0.000)	(0.000)
PBest					− 0.000***
					(0.000)
PBestDuration					0.004***
					(0.000)
Titles fixed effects	No	No	No	No	Yes
Adjusted R^2	0.000	0.011	0.005	0.012	0.066
N (lists)	6,784,538	6,784,538	6,784,538	6,784,538	6,784,538
N (games)	13,952,804	13,952,804	13,952,804	13,952,804	13,952,804
Part II: OLS coefficient (std. dev.) of different Female × Reach interactions					
Female × Reach10			0.326***	0.450***	0.404***
			(0.007)	(0.010)	(0.010)
Female × Reach20			0.363***	0.484***	0.428***
			(0.009)	(0.009)	(0.009)
Female × Reach50			0.377***	0.497***	0.431***
			(0.008)	(0.008)	(0.008)
Female × Reach100			0.403***	0.517***	0.470***
			(0.010)	(0.010)	(0.009)

Specifications in Part II correspond to specifications columns 3, 4, and 5 in Part I, just varying the variable Reach to consider different distances from the personal best. Only coefficient estimates (std. dev.) for the interaction term are reported. *, **, *** denote significance at 5%, 1%, and 0.1%.

TABLE 3.—PERFORMANCE AFTER SETTING A NEW PERSONAL BEST

	Performance relative to rating: $(p_{t+1,t+2} - r_{t+1})/r_{t+1}$				
Dependent variable:	(1)	(2)	(3)	(4)	(5)
Constant	1.314***	24.750***	− 0.678**	24.480***	100.20***
	(0.098)	(0.258)	(0.122)	(0.263)	(1.000)
Female	4.905***	− 2.285***	8.182***	− 1.063***	− 1.374***
	(0.279)	(0.283)	(0.354)	(0.362)	(0.364)
Reach10		− 4.083***	3.852***	− 3.855***	− 3.772***
		(0.307)	(0.325)	(0.329)	(0.330)
Female × Reach10			− 6.405***	− 1.581*	− 3.326***
			(0.923)	(0.912)	(0.906)
PBestBroken		− 4.777***	6.564***	− 4.223***	− 6.881***
		(0.234)	(0.230)	(0.253)	(0.257)
Female × PBestBroken			− 10.230***	− 3.745***	− 5.659***
			(0.643)	(0.637)	(0.633)
Age		− 0.682***		− 0.678***	− 0.598***
		(0.006)		(0.006)	(0.006)
PBest					− 0.038***
					(0.000)
PBestDuration					0.003***
					(0.000)
Titles fixed effects	No	No	No	No	Yes
Adjusted R^2	0.000	0.029	0.002	0.029	0.043
N (pairs of lists)	419,798	419,798	419,798	419,798	419,798
N (games)	5,128,463	5,128,463	5,128,463	5,128,463	5,128,463

Standard deviations are in parentheses. *, **, *** denote significance at 5%, 1%, and 0.1%.

interaction with the female gender both continue to have negative and strongly significant effects. Also intuitive, the coefficient on *PBestBroken* is negative and strongly significant. When a new personal best is set, it is harder to perform in a way to set an even higher best rating. Interestingly, the interaction between *Female* and *PBestBroken* is also negative and strongly significant, indicating that women underperform relative to men after setting a new personal best.

TABLE 4.—ACTIVITY AFTER SETTING A NEW PERSONAL BEST

	Activity growth rate				
Dependent variable:	(1)	(2)	(3)	(4)	(5)
Constant	-0.553^{***}	0.617^{***}	0.558^{***}	0.615^{***}	0.849^{***}
	(0.002)	(0.006)	(0.002)	(0.006)	(0.023)
Female	-0.015^{**}	-0.023^{**}	0.004	-0.017^{*}	-0.028^{***}
	(0.007)	(0.007)	(0.009)	(0.008)	(0.009)
Reach10		0.209^{***}	0.219^{***}	0.202^{***}	0.210^{***}
		(0.007)	(0.007)	(0.007)	(0.007)
Female × Reach10			0.053^{***}	0.063^{***}	0.056^{**}
			(0.023)	(0.023)	(0.023)
PBestBroken		-0.191^{***}	-0.159^{***}	-0.184^{***}	-0.185^{***}
		(0.006)	(0.005)	(0.006)	(0.006)
Female × PBestBroken			-0.068^{***}	-0.053^{***}	-0.060^{***}
			(0.017)	(0.017)	(0.017)
Age		-0.001^{***}		-0.001^{***}	-0.001^{***}
		(0.000)		(0.000)	(0.000)
PBest					-0.000^{***}
					(0.000)
PBestDuration					0.000^{***}
					(0.000)
Titles fixed effects	No	No	No	No	Yes
Adjusted R^2	0.000	0.003	0.003	0.003	0.003
N (pairs of lists)	812,453	812,453	812,453	812,453	812,453
N (games)	6,663,152	6,663,152	6,663,152	6,663,152	6,663,152

Standard deviations are in parentheses. *, **, *** denote significance at 5%, 1%, and 0.1%.

This effect is also quite sizable in all the specifications. For example, for a player with an ELO rating of 2,000, it translates into a decrease in performance of $2 \times 5.66 = 11.32$ points. In appendix A we have also considered, as in the previous tables, different *ReachX* distances and found that their interaction with *Female* remains significant at conventional levels.[24] Importantly, the interaction between *Female* and *PBestBroken* remains quite significant as well, and it increases in size as different *ReachX* distances farther away from the personal best are considered.[25]

The dependent variable in table 4 is the growth rate in playing activity, $(n_{t+1,t+2} - n_{t,t+1})/n_{t,t+1}$. As in table 2, we find that when the personal best has not been broken, having a rating r_{t+1} within reach of 10 rating points from the personal best and its interaction with *Female* both have a positive impact on the rate of playing activity. When the personal best is within reach, women become relatively more active than men. In appendix A we find that this result is present even within a 100 rating point distance, which as indicated earlier is quite large by chess standards. Interestingly enough, both *PBestBroken* and its interaction with *Female* have negative significant coefficients. This means that the rate of playing activity drops immediately after breaking the personal best and, importantly, it drops significantly more for women than for men.[26] This effect is always present and appears to be quite stable in magnitude at around 6% in the last three specifications regardless of the independent variables included. Finally, with respect to the rest of the variables, we find similar effects to those in table 2.

Overall, we take these findings as showing differential responses by gender in terms of both performance and changes in activity levels after setting a new personal best.

VI. Discussion

The results in the previous section are consistent with the existence of behavioral gender differences in dependence around a reference point (personal best). In the extensive margin, women increase their effort more than men before setting a new personal best but exert less effort after doing so. In the intensive margin, women underperform relative to men both to set a new personal best and after doing so.

A number of extensions and refinements are possible, in addition to those discussed previously. For instance, because women in chess may be different from men in various respects, we have followed the "equivalence criteria" across genders used by the FIDE and study in appendix B a more comparable subset of men and women for our main age group.[27] Although the evidence shows some differences across the different subsets of subjects considered that can

[24] Interestingly, the size (in absolute value) tends to decrease up to a within a 20–40 rating point distance as in table 1 and to increase after that.

[25] This interaction effect loses most of its significance for the younger group (AgeU20) and is greater in magnitude in the main age group (Age20–60), for whom the coefficient always decreases in absolute value with the distance to the personal best.

[26] In appendix A we vary the distance *ReachX* and find that the interaction between *Female* and *PBestBroken* decreases as we consider greater distances, except in the main Age20–60 group for whom this interaction term is always similar in size and strongly significant.

[27] In particular, we consider ratings in the interval [2000, 2600] for men and [1800, 2400] for women. This balancing or comparability criterion reflects the 200 rating point difference in the FIDE regulations to award the *Grandmaster* and *International Master* titles versus the corresponding versions *Women Grandmaster* and *Women International Master* titles. These different intervals are intended to cover similar interpercentile ranges within the respective populations. In addition, we also consider

be of interest in their own right, the basic empirical findings are robust and essentially remain unchanged. Similarly, in this appendix we also provide evidence from matching estimators where we compare women and men of the same age, same personal best, and same personal best duration for this group. The results for the relevant interactions of interest do not vary much and continue to be robust.

We have also implemented a number of additional robustness checks. We briefly discuss three of them next.[28] First, in principle, it is not impossible to think that different circumstances may exist for men and women that could themselves vary when close and far from the personal bests. Although it seems unlikely, this might affect the extensive margin. For instance, the availability of tournaments in which to participate may not be equal between men and women when they are 15 points near a personal best and 100 points. Tournament availability for players is difficult to control for because the location of players and geographical distance to tournaments is not available in the FIDE data set. However, it is possible to check whether availability is a relevant aspect using the fact that the density of tournaments is not uniform during the year. Tournaments are more concentrated during the summer months (July and August), which is when players have more time, and hence when there is more, cheaper, or at the very least different access to tournaments than in other parts of the year. If differences in availability are a relevant determinant of differences in participation, we should see that the effects are different in different parts of the year.[29] To study this, we have implemented the same regression specifications for each part of the year (summer and non-summer) separately. The results show no significant differences, which is consistent with the hypothesis of no differential impact of tournament availability for men and women when close and far from the personal bests.

A second robustness test concerns the frequency with which ratings lists are updated by the FIDE, which nowadays is every month. The reason is that a player may beat temporarily his or her personal best during a month, but this need not be reflected in the official list published after the month ends (if after playing more games he or she is back below his initial personal best before the end of the month). That is, both the "true" and the "official" personal bests may matter in a player's mind, but they need not always coincide. Although, in principle, it is unclear whether this aspect could be relevant in practice, it is something worth examining.[30] We can then check whether the results are robust by studying changes in how frequently lists are updated from monthly to quarterly. If official lists were published, say, once a decade, then the official personal best would likely be irrelevant as a reference point (the only thing that might matter would be the unofficial personal best, if known). That is, if the results are robust, we should find that the effects are present at all the frequencies in the data and that, if anything, there are stronger performance effects the greater the updating frequency of lists. This is exactly what we find. We use the exogenous variation in the frequency with which rating lists are updated by the FIDE (see footnote 8) and study the results at monthly frequencies (July 2012 to date) and at lower frequencies (January 2000 to June 2012). The results confirm that the effects we have documented are highly significant at all frequencies, and that there are stronger performance effects when the updating frequency is higher.[31]

Third, in the FIDE data set, the probability of playing against a man slowly increases as players' ratings increase. Recent research in psychology suggests the possibility of an opponent gender (OG) effect in chess whereby women may underperform when playing against men (compared to a woman of the same ELO rating).[32] Although for the ELO range that we study (increases within a 10, 20, or 50 ELO point distance from the personal best), this increase is statistically no different from zero, it is worth examining this aspect. Say that in addition to the detrimental distance-to-personal best effect (DPB) in women's performance that we have documented that we find an OG effect. We can exploit the fact that the OG effect is independent of the distance from the personal best, whereas a DPB should tend to disappear as the distance grows larger. We have taken advantage of this asymmetry and studied what happens far from the personal best, in particular at greater than distances of 50 ELO points. More precisely, we have examined players' performances in the main age group 20–60 when their "personal best minus 50 points" is within reach. Since reaching this new "maximum" has nothing to do with breaking a personal best, there is no room for a DPB effect. On the other hand, the OG effect should still be present, if it exists. The results show that the gender effects are not significantly different from zero; that is, they are consistent with the hypothesis that the OG effect is nil and in practice not relevant.

Finally, a brief word on the generalizability. A natural question is how much the results are likely to generalize to other settings. Although speculative, some discussion is warranted. We have a setting where subjects perform a cognitive task, in a strictly competitive environment, and one

specific subsets according to different levels of experience, different maximum duration of the personal best (in terms of games and months), and different distances to the personal best.

[28] We are grateful to two anonymous referees for suggesting them. They are also included in appendix B. In general, it may be noted that in most regressions the adjusted R^2 is small. This is to be expected given the massive amount of data we have and their variability.

[29] See Barnanchon, Rathelot, and Roulet (2021) for a study relating gender differences in willingness to commute to the gender wage gap.

[30] First, even the players themselves may not know when they have beaten their personal best (other than in the official lists, of course), as they may not have the tools to compute their rating on a game-by-game basis (not even today, much less one or two decades ago). Second, beating one's personal best in a way that is not "official" cannot easily be "proven" to others, which players probably care about.

[31] In terms of participation, the results are similar across frequencies after breaking the personal best, and slightly stronger before that at lower frequencies. This suggests the possibility of some substitution from the intensive to the extensive margin.

[32] These findings are somewhat inconclusive though; see Stafford (2018) and Smerdon et al. (2020).

where women are a relatively small percentage of subjects. In principle, settings that share some or all of these characteristics (cognition, competition, and gender-skewed environments) would appear to be candidates for observing a generalization of the results. Interestingly, these are some of the settings that receive substantial attention in the social sciences and in the media. As occupational sorting and segregation by gender remains high in labor markets, there is important research trying to understand gender performance and integration in male-dominated work environments. The role of preferences, stereotypes, and social norms also has far-reaching implications for policies aimed at integrating the workplace.[33] Similarly, a substantial amount of research has tried to understand the mechanisms underlying how gender affects achievement in schools and in the workplace. A prominent body of research, for instance, tries to explain why fewer women than men pursue careers in STEM subjects. Our results suggest that behavior around personal bests represents a new promising avenue for future research on these and related questions.

Future research should also include new theoretical developments. In terms of theoretical frameworks, our results indicate that gender emerges as a determinant of reference-dependent behavior. Therefore, they readily suggest including gender in models of reference dependence, a conclusion that may in fact be taken as a main contribution of our study. Although the goal of our study is not theoretical, the effort model of loss aversion in Anderson and Green (2018), for example, can be made consistent with observed patterns if the performance and costs functions are suitably extended to allow for gender dependence. Similarly, the models in Alaoui and Penta (2016, 2021) on cost-benefit and endogenous depth of reasoning could also be extended along the gender dimension.

VII. Conclusion

Reference dependence is a fundamental principle of human behavior that captures the comparative nature of human feelings and perceptions. In spite of its importance, and the existence of a large literature on the broad applicability of various forms of reference dependence, little is known about how this determinant of behavior may vary across demographic characteristics. In particular, little is known about a reference point that is part of every human subject hedonic experience: his or her personal best. Motivated by an influential and voluminous literature that studies gender, but that is silent about reference points, our goal has been to provide a first study linking these bodies of scientific inquiry.

We have studied a real-life cognitive task in a strictly competitive setting with subjects who are experts and officially ranked, and who perform under high stakes. In this setting, performance and participation can be clearly and precisely documented, and we can take advantage of the availability of the universe of official competitions. Our results support the hypothesis that the comparative nature of human behavior is different across genders. Although many differences in preferences and attitudes across genders have been previously documented in the literature, these findings open up new avenues of future research. We hope they will motivate researchers to undertake future empirical and theoretical study linking gender and reference dependence, using different reference points, different competitive and cooperative settings, and different incentives, tasks, subjects, and other characteristics.

[33] Dahl, Kotsadam, and Rooth (2021), for instance, examine whether integrating men and women in a male-dominated environment (the military) can change men's attitudes about mixed-gender productivity (mostly in noncognitive tasks), gender roles, and gender identity.

REFERENCES

Alaoui, L., and A. Penta, "Endogenous Depth of Reasoning," *Review of Economic Studies* 83 (2016), 1297–1333. 10.1093/restud/rdv052

——— "Cost-Benefit Analysis in Reasoning," *Journal of Political Economy*, 130 (2021), 881–925. 10.1086/718378

Allen, E. J., P. M. Dechow, D. G. Pope, and G. Wu, "Reference-Dependent Preferences: Evidence from Marathon Runners," *Management Science* 63 (2017), 1657–1672. 10.1287/mnsc.2015.2417

Anderson, A., and E. A. Green, "Personal Bests as Reference Points," *Proceedings of the National Academy of Sciences* 115 (2018), 1772–1776. 10.1073/pnas.1706530115

Azmat, G., and B. Petrongolo, "Gender and the Labor Market: What Have We Learned from Field and Lab Experiments?" *Labour Economics* 30 (2014), 32–40. 10.1016/j.labeco.2014.06.005

Barnanchon, L., T. R. Rathelot, and A. Roulet, "Gender Differences in Job Search: Trading Off Commute against Wage," *Quarterly Journal of Economics* 136 (2021), 381–426. 10.1093/qje/qjaa033

Becker, G. S., *Accounting for Tastes* (Cambridge, MA: Harvard University Press, 1996).

Becker, G. S., and K. M. Murphy, *Social Economics: Market Behavior in a Social Environment* (Cambridge, MA: Belknap Press of Harvard University Press, 2000).

Benabou, R., and J. Tirole, "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics* 117 (2002), 871–915. 10.1162/003355302760193913

——— "Intrinsic and Extrinsic Motivation," *Review of Economic Studies* 70 (2003), 489–520. 10.1111/1467-937X.00253

——— "Willpower and Personal Rules," *Journal of Political Economy* 112 (2004), 848–887. 10.1086/421167

Bernheim, D., S. DellaVigna, and D. Laibson (eds.), *Handbook of Behavioral Economics—Foundations and Applications* (Amsterdam: North-Holland, 2018).

Bordalo, P., N. Gennaioli, and A. Shleifer, "Salience Theory of Choice under Risk," *Quarterly Journal of Economics* 127 (2012), 1243–1285. 10.1093/qje/qjs018

——— "Salience and Consumer Choice," *Journal of Political Economy* 121 (2013), 803–843. 10.1086/673885

——— "Memory, Attention and Choice," *Quarterly Journal of Economics* 135 (2020), 1399–1442. 10.1093/qje/qjaa007

Bowles, S., "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions," *Journal of Economic Literature* 36 (1998), 75–111.

Burdina, M., R. S. Hiller, and N. E. Metz, "Goal Attainability and Performance: Evidence from Boston Marathon Qualifying Standards," *Journal of Economic Psychology* 58 (2017), 77–88. 10.1016/j.joep.2017.01.001

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

Cunha, F., J. J. Heckman, and S. M. Schennach, "Estimating the Technology of Cognitive and Non-cognitive Skill Formation," *Econometrica* 78 (2010), 883–931.

Dahl, G. B., A. Kotsadam, and D. Rooth, "Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams," *Quarterly Journal of Economics* 136 (2021), 987–1030. 10.1093/qje/qjaa047

Dawson, C., and D. de Meza, "Wishful Thinking, Prudent Behavior: The Evolutionary Origin of Optimism, Loss Aversion and Disappointment Aversion," SSRN working paper 3108432 (2018).

Damon, C., D. Gill, M. Rush, and V. Prowse, "Using Goals to Motivate College Students: Theory and Evidence from Field Experiments," this REVIEW 102 (2020), 648–663.

DellaVigna, S., "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature* 47 (2009), 315–372. 10.1257/jel.47.2.315

Dilmaghani, M., "Gender Differences in Performance under Time Constraint: Evidence from Chess Tournaments," *Journal of Behavioral and Experimental Economics* 89 (2020), 101505. 10.1016/j.socec.2019.101505

Ehrenberg, R. G., and M. L. Bognanno, "Do Tournaments Have Incentive Effects?" *Journal of Political Economy* 98 (1990), 1307–1324. 10.1086/261736

Fehr, E., and K. Hoff, "Tastes, Castes, and Culture: The Influence of Society on Preferences," *Economic Journal* 121 (2011), 396–412. 10.1111/j.1468-0297.2011.02478.x

FIDE, *FIDE Rating Regulations. FIDE Handbook (Chapter B.02)* (Lausanne: World Chess Federation, 2017).

Garicano, L., I. Palacios-Huerta, and C. Prendergast, "Favoritism under Social Pressure," this REVIEW 87 (2005), 208–216.

Gill, D., and V. Prowse, "Cognitive Ability, Character Skills, and Learning to Play Equilibrium: A Level-*k* Analysis," *Journal of Political Economy* 126 (2016), 1619–1676.

Goldin, C., "The Quiet Revolution That Transformed Women's Employment, Education, and Family," *American Economic Review* 96 (2006), 1–21. 10.1257/000282806777212350

——— "A Grand Gender Convergence: Its Last Chapter," *American Economic Review* 104 (2014), 1091–1119. 10.1257/aer.104.4.1091

González-Díaz, J., O. Gossner, and B. Rogers, "Performing Best When It Matters Most: Evidence from Professional Tennis," *Journal of Economic Behavior & Organization* 84 (2012), 767–781.

González-Díaz, J., and I. Palacios-Huerta, "Cognitive Performance in Competitive Environments: Evidence from a Natural Experiment," *Journal of Public Economics* 139 (2016), 40–52.

Harding, M., and A. Hsiaw, "Goal Setting and Energy Conservation," *Journal of Economic Behavior and Organization* 107 (2014), 209–227. 10.1016/j.jebo.2014.04.012

Heath, C., R. P. Larrick, and G. Wu, "Goals as Reference Points," *Cognitive Psychology* 38 (1999), 79–109. 10.1006/cogp.1998.0708

Heckman, J. J., and T. Kautz, "Fostering and Measuring Skills: Interventions That Improve Character and Cognition" (pp. 341–430), in *The Myth of Achievement Tests: The GED and the Role of Character in American Life* (Chicago: University of Chicago Press, 2014).

Koch, A. K., and J. Nafziger, "Goals and Bracketing under Mental Accounting," *Journal of Economic Theory* 162 (2016), 305–351. 10.1016/j.jet.2016.01.001

Köszegi, B., and M. Rabin, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics* 121 (2006), 1133–1165.

——— "Reference-Dependent Risk Attitudes," *American Economic Review* 97 (2007), 1047–1073.

——— "Reference-Dependent Consumption Plans," *American Economic Review* 99 (2009), 909–936.

Markle, A., G. Wu, R. White, and A. Sackett, "Goals as Reference Points in Marathon Running: A Novel Test of Reference Dependence," *Journal of Risk and Uncertainty* 56 (2018), 19–50. 10.1007/s11166-018-9271-9

Niederle, M., and L. Vesterlund, "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics* 122 (2007), 1067–1101. 10.1162/qjec.122.3.1067

O'Donoghue, T., and C. Sprenger, "Reference-Dependent Preferences" (pp. 1–77), in *Handbook of Behavioral Economics—Foundations and Applications* (Amsterdam: Elsevier, 2018).

Olivetti, C., and B. Petrongolo, "The Evolution of Gender Gaps in Industrialized Countries," *Annual Review of Economics* 8 (2016), 405–434. 10.1146/annurev-economics-080614-115329

Palacios-Huerta, I., "Professionals Play Minimax," *Review of Economic Studies* 70 (2003), 395–415. 10.1111/1467-937X.00249

——— *Beautiful Game Theory* (Princeton, NJ: Princeton University Press, 2014).

——— "Nandi Female Husbands," SSRN working paper 3899091 (2021).

Palacios-Huerta, I., and T. J. Santos, "A Theory of Markets, Institutions, and Endogenous Preferences," *Journal of Public Economics* 88 (2004), 601–627. 10.1016/S0047-2727(02)00162-7

Palacios-Huerta, I., and O. Volij, "Field Centipedes," *American Economic Review* 99 (2009), 1619–1635. 10.1257/aer.99.4.1619

Pope, D., and U. Simonsohn, "Round Numbers as Goals: Evidence from Baseball, SAT Takers, and the Lab," *Psychological Science* 22 (2011), 71–79. 10.1177/0956797610391098

Pope, D. G., and M. E. Schweitzer, "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes," *American Economic Review* 101 (2011), 129–157. 10.1257/aer.101.1.129

Proto, E., A. Rustichini, and A. Sofianos, "Intelligence, Personality, and Gains from Cooperation in Repeated Interactions," *Journal of Political Economy* 127 (2019), 1351–1390. 10.1086/701355

Rayo, L., and G. S. Becker, "Evolutionary Efficiency and Happiness," *Journal of Political Economy* 115 (2007), 302–337. 10.1086/516737

Romer, D., "Do Firms Maximize? Evidence from Professional Football," *Journal of Political Economy* 114 (2006), 340–365. 10.1086/501171

Smerdon, D., H. Hu, A. McLennan, W. von Hippel, and S. Albrecht, "Female Chess Players Show Typical Stereotype-Threat Effects: Commentary on Stafford," *Psychological Science* 31 (2020), 756–759. 10.1177/0956797620924051

Smithers, S., "Goals, Motivation and Gender," *Economics Letters* 131 (2015), 75–77. 10.1016/j.econlet.2015.03.030

Stafford, T., "Female Chess Players Outperform Expectations When Playing Men," *Psychological Science* 29 (2018), 429–436. 10.1177/0956797617736887

Szymanski, S., "A Market Test for Discrimination in the English Professional Soccer Leagues," *Journal of Political Economy* 108 (2000), 590–603. 10.1086/262130

Williams, E. F., and T. Gilovich, "The Better-than-My-Average Effect: The Relative Impact of Peak and Average Performances in Assessments of Self and Others," *Journal of Experimental Social Psychology* 48 (2012), 556–561. 10.1016/j.jesp.2011.11.010