Judgment and Decision Making, Vol. 13, No. 5, September 2018, pp. 413–427

Randomization and serial dependence in professional tennis matches: Do strategic considerations, player rankings and match characteristics matter?

Leonidas Spiliopoulos*

Abstract

In many sports contests, the equilibrium requires players to randomize across repeated rounds, i.e., exhibit no temporal predictability. Such sports data present a window into the (in)efficiency of random sequence generation in a natural competitive environment, where the decision makers (tennis players) are both highly experienced and incentivized compared to laboratory studies. I resolve a long-standing debate about whether professional players' tennis serve directions are serially independent (Hsu, Huang & Tang, 2007) or not (Walker & Wooders, 2001) using a new dataset that is two orders of magnitude larger than those studies. I examine both between- and within-player determinants of the degree of serial (in)dependence. Evidence of the existence of significant serial dependence across serves is presented, even among players ranked Number 1 in the world. Furthermore, significant heterogeneity was found with respect to the strength of serial dependence and also its sign. A novel finding is that Number 1 and Number 2 ranked players tend to under-alternate on average, whereas in line with previous findings, the lower-ranked the players, the greater their tendency to over-alternate. Within-player analyses show that high-ranked players do not condition their randomization behavior on their opponent's ranking. However, the under-alternation of top players would be consistent with a best-response to beliefs that the population of opponents over-alternates on average. Finally, the degree of observed serial dependence is not systematically related to other match variables proxying for match difficulty, fatigue, and psychological pressure.

Keywords: randomization, mixed strategy Nash equilibrium, minimax, tennis, sports data analytics

1 Introduction

The production and perception of randomness has a long research history in cognitive psychology (see Nickerson, 2002, for an overview) and rightly so. The perception or judgment of randomness is a core human competency (see Oskarsson et al., 2009, for a review). There is ample evidence that humans are capable of learning patterns (both implicitly and explicitly) in sequences of events (Clegg et al., 1998; Remillard & Clark, 2001). Our ability to discover the correlations (e.g., Kareev, 1995; Kareev et al., 1997; Kareev, 2000) arising from the causal relationships in our environment allow us to adapt to and exploit the environmental structure. With respect to the production or generation of random behavior, subjects in laboratory tasks (without strategic interactions) are typically inefficient at creating serially uncorrelated sequences. Subjects tend to produce over-alternating sequences (with too many runs) and regress towards the representative frequencies of the distribution they are emulating (Kahneman & Tversky, 1972; Bar-Hillel & Wagenaar, 1991; Rapoport & Budescu, 1997). Explanations of these deviations in random generation range from cognitive bounds such as short-term memory (Kareev, 1992, 1995, 2000) and the complexity (or difficulty of encoding) of sequences (Falk & Konold, 1997) to the statistical properties of small samples of random behavior (Kareev et al., 1997; Sun & Wang, 2010, 2011), or the interaction of both (Hahn & Warren, 2009; Farmer et al., 2017; Warren et al., 2018).

Although the *judgment* of randomness is typically applicable to interactions with nature or individual decision making, the *production* or *generation* of random behavior is naturally most relevant to strategic interactions with other decision makers in our environment, i.e., in strategic games. In contrast to the above studies that investigate random sequence generation in individual decision-making tasks, Rapoport & Budescu (1992) and Budescu & Rapoport (1994) used laboratory games where it is optimal to be unpredictable. While they found similar qualitative deviations in randomization behavior for both individual and strategic decision-making, Budescu & Rapoport (1994) show that people are more efficient randomizers in the latter. Random behavior is called for in strategic interactions of conflict or competition, where one player's gain is another's loss and being unpredictable

The author would like to thank Andreas Ortmann and John Wooders for constructive feedback and gratefully acknowledges financial support for this project from the Alexander von Humboldt Foundation (Humboldt Research Fellowship for Experienced Researchers).

Copyright: © 2018. The authors license this article under the terms of the Creative Commons Attribution 3.0 License.

*Max Planck Institute for Human Development, Center for Adaptive Rationality, Lentzeallee 94 Berlin, 14195, Germany. Email: spiliopoulos@mpib-berlin.mpg.de

is beneficial. Such games have an equilibrium in mixed-strategies, where a player chooses to randomize over the actions at his/her disposal, rather than play one of them with certainty. Situations where mixed strategies are relevant include bluffing in poker, penalty shootouts in soccer, and serve directions in tennis, which is the environment that I will study here. In repeated games, the normative prediction is that the action chosen in a round should be independent of chosen actions in the previous round, i.e., players should be randomizing perfectly. Otherwise, a player could learn the dependencies or patterns in the opponent's behavior and exploit them appropriately (Spiliopoulos, 2012, 2013a,b, 2018; Ioannou & Romero, 2014).

Field data from competitive sports are particularly useful, combining the benefits of a real-world domain where randomization is important (guaranteeing high ecological validity) with a high-level of incentivization and the opportunity for significant learning beyond what is feasible in the laboratory. The existing literature using field data has been primarily conducted by game theorists and economists rather than cognitive psychologists, despite its obvious relationship to pioneering work by psychologists on randomization. In this paper, I analyse a large dataset of tennis serves with the goal of resolving an open debate on whether professional players deviate from efficient randomization in their serve direction (Walker & Wooders, 2001) or not (Hsu et al., 2007). Furthermore, I extend the existing literature by exploiting the large number of within-player observations to examine whether the degree of randomization depends on a player's own rank and the rank of the opponent, experience, the round of the match (e.g., final, semi- or quarter-final), and the difficulty and length of the match. These analyses are related to existing laboratory studies investigating the impact of learning, feedback and other variables on the efficiency of randomization. Specifically, Lopes & Oden (1987) concluded that statistically sophisticated subjects performed better than average subjects, although they exhibited the same qualitative misperceptions of randomness. Regarding whether feedback induces better randomization, the evidence thus far is mixed. Feedback has been found to improve the identification of non-random sequences (Zhao et al., 2014) and generation of random sequences (Neuringer, 1986); however, Budescu (1987) did not find a significant effect of feedback. Of course, the degree of learning that can occur in the laboratory is limited by practical ceilings on the amount of exposure and the incentives to perform well. The field data from highly-paid and competitive tennis tournaments addresses both of these limitations and permits the investigation of other potential mediators.

Before proceeding, I summarise the state of the art in the game theory literature. Recall that the normative solution to repeated games with a stage unique mixed-strategy Nash equilibrium is perfect randomization, i.e., actions must be independent of the prior history of play. One strand of experimental studies tests the equilibrium predictions in the laboratory, finding significant deviations from the equilibrium predictions (Bloomfield, 1994; Brown & Rosenthal, 1990; Chiappori et al., 2002; Ochs, 1995; Rapoport & Budescu, 1997; O'Neill, 1987; Levitt et al., 2010; Wooders, 2010; Palacios-Huerta & Volij, 2008; Okano, 2013; Shachat, 2002). Experience can reduce the magnitude of these deviations, however this is conditional on features of the game — see Ochs (1995); Roth & Erev (1995); Erev & Roth (1998); Binmore et al. (2001); Nyarko & Schotter (2002). Another finding is that experience from the field does not transfer well to the laboratory for new tasks. Despite initial claims that professionals, to a large degree, transfer their experience to new tasks in the laboratory (Palacios-Huerta & Volij, 2008), later studies have not found evidence of this effect (Levitt et al., 2010; Wooders, 2010; Van Essen & Wooders, 2015). Finally, subjects exploit both deviations from the equilibrium marginal distributions (Shachat & Swarthout, 2004) and deviations from serially independent or random play, in ways that can be explained by learning models capable of detecting temporal patterns (Spiliopoulos, 2012, 2013a,b, 2018; Ioannou & Romero, 2014).

Another strand of research utilizes field data from competitive sports. The first paper to examine the optimality of tennis serves in the field is Walker & Wooders (2001) — see also the comment by Hsu et al. (2007) (I refer to these two studies as WW and HHT respectively). Both studies concluded that mixing proportions were not statistically different from the equilibrium; however, while the former concluded that significant deviations existed from the theoretical prediction of serial independence, the latter concluded the opposite. The predictions of minimax play in the field have also been tested in other sports, such as soccer and the NFL.[1] To summarize, the majority of studies confirm equilibrium behavior in terms of mixing proportions, whereas the findings regarding serial independence are mixed. Of these different sports, tennis allows for the most powerful tests of minimax behavior for individual players rather than a population of players. In soccer, since players rarely make penalty shots, the data afford low statistical power to reject the null hypothesis of equilibrium behavior at the individual level. Also, because there exist large intervals between

[1] In soccer, Chiappori et al. (2002) conclude that the mixing proportions of penalty kicks are in accordance with theoretical predictions; a re-analysis by Coloma (2007) of their data directly testing mixing proportions and new data by Buzzacchi & Pedrini (2014) confirm this finding. Similarly, Palacios-Huerta (2003) found that mixing proportions are in line with the equilibrium prediction, and that serial independence across penalties could not be rejected. Dohmen & Sonnabend (2016) conclude the same on both counts. Kovash & Levitt (2009) find significant deviations from the theory in baseball pitches and NFL plays for both mixing proportions and serial independence (on average over-alternation is more common). Similarly, Emara et al. (2014) conclude that there exists a significant bias towards over-alternating in NFL plays. On the other hand, McGarrity & Linnen (2010) find that play in the NFL is not significantly different from the equilibrium with respect to mixing proportions and serial independence.

a player's consecutive penalties, this could encourage equilibrium behavior by inducing memory-less behavior, which may be conducive to the generation of serially independence sequences. In the NFL, different players are involved in each play and strategies are called by the coach; hence, tests of equilibrium behavior are essentially a test not of an individual, but a joint test of the behavior of the coach and a group of players.

This study is most similar to WW and HHT, but uses a new tennis serve dataset that is two orders of magnitude larger than those in existing published studies. I will specifically address the conflicting findings regarding serial (in)dependence in WW and HHT, which remains an important open issue. While WW do not reject serial independence of serves, HHT find evidence of statistically significant correlation across serves. I extend the work in WW and HHT in three directions. First, by including analyses of behavior relative to the player's (own) ranking, i.e., examining whether more highly ranked players conform more closely to equilibrium predictions. A working paper by Gauriot et al. (2016), henceforth GPW, uses another large dataset from another source to examine minimax behavior in tennis and its relationship to player ranking. Note, the latter manuscript also investigates the equilibrium prediction that winning rates for left and right serves are equal. While there is some overlap between our manuscripts in terms of the hypotheses tested, they are largely complementary.

The following hypotheses based on individual player-level analyses (rather than only population analyses) differentiate my work from WW, HHT and GPW. The first hypothesis regards whether players strategically condition their behavior on the ranking of their opponent. Players capable of using the equilibrium strategy — but consciously choosing not to play accordingly — may in fact be rational if they hold *correct* beliefs that their opponent will not choose the normative solution (Plott, 1996). Consider the case where low ranked players are imperfect randomizers. If high-ranked players are sophisticated in the sense of correctly predicting low-ranked players' deviations from the equilibrium, then rationality dictates that they exploit this. Of course, this would lead to non-equilibrium behavior by the highly-ranked players, which however, would indicate rational behavior given their opponent's type. The second set of hypotheses regard whether players condition on, or are affected by, match characteristics such as: a) the tournament round of the match (e.g., whether it is a final, semi-final etc.), which would factor in the effects of stress and the probability of winning the tournament prize given the tournament's progression, b) the difficulty of the match (e.g., how close the score is), and the number of points played in a match (a proxy for fatigue and difficulty). To the best of my knowledge this is the first field study in tennis to address all of these additional questions.

TABLE 1: The specification of a point game in terms of the server winning probabilities, π_{a_s,a_r}

		Receiver L	*Receiver* R
Server L		$\pi_{L,L}$	$\pi_{L,R}$
Server R		$\pi_{R,L}$	$\pi_{R,R}$

2 Modeling tennis serves

I briefly describe the tennis serve model introduced by WW and adopted by HHT (see WW for more details). Tennis serves alternate in terms of the area of the court (box) within which each serve must land to be valid (i.e., not declared as a fault), referred to as *deuce* and *ad* courts. The collections of points served by a player in each of these two courts are referred to as either ad or deuce point-games. For example, all points in a match where a specific player's serve was directed to the ad box are referred to as that player's ad point-game. Since there are two players and two boxes, each match has four point-games. Each point-game in the match is modeled as a 2×2 normal form game with action spaces left (L) and right (R) for both the server s and the receiver r — see Table 1. The payoffs of this game are equivalent to the probabilities π_{a_s,a_r} of winning each point-game for the action profile a_s, a_r — consequently, this game is constant-sum. The probabilities of winning differ conditional on whether serves are made to the ad or deuce court due to differences in serving and returning abilities; this is the reason why we must distinguish between point-games. Walker et al. (2011) show that tennis belongs to the class of Binary Markov games, which possess the property that the equilibrium play for every point in the match can be solved independently of all other past points and outcomes in the match. That is, the equilibrium of the match corresponds to equilibrium play in each point of a specific point-game. Every point-game played has a unique mixed strategy Nash equilibrium under the assumptions that $\pi_{L,L} < \pi_{R,L}$, $\pi_{R,R} < \pi_{L,R}$, $\pi_{L,L} < \pi_{L,R}$ and $\pi_{R,R} < \pi_{R,L}$.[2]

3 Data

The data originate from the crowd-sourced Match Charting Project accessible at http://www.tennisabstract.com/charting/meta.html, which compiles tennis match statistics.[3]

[2]These inequalities follow from the reasonable assumption that the server is more likely to win a point if the direction of the serve and the direction anticipated by the receiver are mismatched.

[3]The data at the Match Charting Project are updated often with new statistics as volunteers upload information from more tennis matches. The dataset used for the analysis was downloaded on 6/6/2016. Preliminary analyses performed using snapshots of this dataset at various points in time in 2015 (i.e., comprised of subsets of the final dataset) led to similar conclusions. Further information on the Match Charting Project can be found at: http://www.tennisabstract.com/charting/meta.html and https://github.com/

The dataset covers 391 male players whose career-high ranking ranged from Number 1 to Number 2076 (mean = 123, median = 71) in the world, and includes 1,093 matches from 1975 to 2016.[4] In total, I analyze the data from 143,743 serves resulting from 4,372 point-games. This is two orders of magnitude larger than prior published studies of tennis serves: 3,026 serves from ten matches in WW and 2,490 serves from ten mens matches in HHT.[5] The mean (median) number of matches per player is 5.6 (2); for top players who are more likely to participate in these tournaments the dataset holds significantly more matches, e.g., the maximum is for Federer (160 matches), followed by Nadal (142), Djokovic (126) and Murray (68). For these players there are 12413, 8708, 8984, and 4935 serve observations respectively. This amount of data permits much more powerful tests of the mixing behavior of athletes than previous studies. Furthermore, hypothesis testing targeted at the top players provides the best chance of observing equilibrium behavior, as these players are the most capable and the most highly incentivized to pursue optimal behavior.

In the dataset, tennis serve directions were encoded as either "4", "5", or "6", corresponding to left, center and right respectively. I found 21,159 cases, where other symbols were used in the encoding. This may be either due to data-entry error, or because the data-coders were uncertain how to categorize the serve direction. These cases were not included in the analysis, as is the case for serves in the direction of the center — the latter is standard practice in the literature, i.e., prior studies analyzed only the left and right serve directions. The complete dataset was compiled by merging point-by-point data files with player-ranking data files ranging from 22/12/1980 to 1/2/2016. I use both the career-high and current ranking (at the time of the match) in the analyses; in some cases the former may be more appropriate as players' current rankings may misleadingly fluctuate wildly due to injuries.[6] The following notation is used throughout. Let i index the players, pg index the point-games from all players' matches, and pg_i index the point games for a player i.

JeffSackmann/tennis_MatchChartingProject. The data are available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0) https://creativecommons.org/licenses/by-nc-sa/4.0/.

[4]The data from one match were excluded as the column titles clearly did not match the data entries (Match-ID=20130810, Aptos Tournament, Semi-final between Klahn and Donskoy on 10/08/13).

[5]Magnus & Klaassen (1999); Klaassen & Magnus (2009) used a large dataset of 59,466 observations from Wimbledon, but do not report information about serve directions.

[6]Seven missing values for player career-high rankings were replaced with the rankings as recorded at the ATP website http://www.atpworldtour.com/en/rankings/singles.

4 Results

The serial independence of tennis serve directions is tested at two different levels of aggregation: the (dis-aggregated) point-game level and the player level (aggregating over the point-games of each player). WW and HHT tested serial dependence using the distribution of *point-game* statistics since they did not have not enough observations per player. Testing at the player-level is more desirable because it matches the expected structure of the data, particularly the heterogeneity that may exist between players (based on their ability, experience etc.). Below, I summarize the statistical procedures — details can be found in Appendix A.

Serial dependence for each point-game in the data is examined using the two-sided exact runs test (see Eq. 2). To test whether a *set* of these points-game statistics (either for the whole population of players or for a specific player) are distributed according to the null hypothesis of no serial dependence requires the randomization of the test statistics. These randomized statistics are generated according to Walker & Wooders (2001, p. 1533) — a set of these statistics can then be tested using the standard Kolmogorov-Smirnov test. The individual point-game level test is a KS-test on the distribution of the randomized (exact run) test statistics at the point-game level for all the players — this is the test in WW and HHT. For the player-level analysis it is the Kolmogorov-Smirnov test on the distribution of point-game statistics for each player only. The latter permits the testing of serial independence for each player rather than the set of players.

4.1 Analysis at the point-game level

At the point-game level, the null hypothesis of serially independent serve directions was rejected at the 5% level for 12% of the point-games (9.5% for over-alternating, 2.5% for under-alternating). Controlling for multiple comparisons using the Bonferonni-Holm correction, the hypothesis of no serial correlation is rejected for 172 individual point-games, i.e., 3.9% of the cases (3.7% for over-alternating, 0.2% for under-alternating). Alternatively, following WW and HHT, the randomized KS-test on the distribution of the point-games strongly rejects the null hypothesis of serial independence $\left(K = 0.06, p = 2.2 \times 10^{-14}\right)$. I conclude that the hypothesis of serial independence is rejected, predominantly due to over-alternation of the serve direction. This finding corroborates the conclusions drawn by WW, but not HHT — calculations in Appendix C reveal that, given the samples sizes of these two studies, there would be roughly a 50% chance that two independent studies would arrive at the opposite conclusions. The GPW working paper also rejects serial independence using a large dataset with sufficient power.

TABLE 2: Population averaged marginal and conditional probabilities of serve direction

Marginal			Conditional (Transition matrix)					
	Ad	*Deuce*		*Ad*			*Deuce*	
				L	R		L	R
L	0.537	0.486	L	0.51	0.49	L	0.461	0.539
R	0.463	0.514	R	0.57	0.43	R	0.513	0.487

4.2 Analysis at the player level

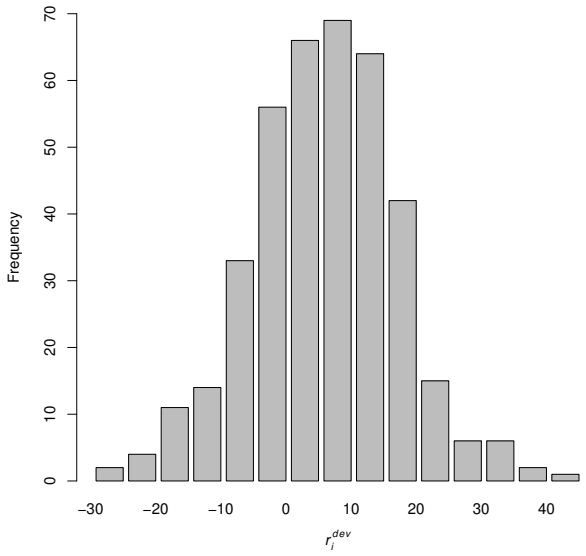
Let the marginal probabilities of each player i serving to the left and to the right be denoted by q_i^L and q_i^R respectively. Recall that these must be separately estimated for the ad and deuce point-games; the subscript related to the point-games is dropped for simplicity. Similarly, the set of conditional probabilities are denoted by $\left\{q_i^{LL}, q_i^{LR}, q_i^{RL}, q_i^{RR}\right\}$, where the first letter of the superscript denotes the serve direction at $t-1$ and the second to the serve direction at t. The conditional probabilities reveal whether players tend to over- or under-alternate. Since these probabilities are conditioned only on the immediately prior serve direction (for the same point-game), they can be represented as the first-order transition matrix of a two-state Markov-chain model:

		Direction at t	
		L	R
Direction at $t-1$	L	q_i^{LL}	q_i^{LR}
	R	q_i^{RL}	q_i^{RR}

Over-alternation implies that $q_i^{LL} < q_i^L$ and $q_i^{RR} < q_i^R$, and under-alternation implies the opposite signs. The maximum likelihood estimates of the marginal and conditional probabilities of serve directions for both point games are presented in Table 2 — these are the averages of the estimates for each point-game in the dataset. As expected, there are differences in the marginal and conditional probabilities for the two point games, arising from differences in the ability to serve and return for the ad and deuce courts. For both the ad and deuce point-games and both serve directions, the population of players exhibit over-alternation on average, as $q^{LL} < q^L$ and $q^{RR} < q^R$ for both point-games — see Table 2.

Aside from the conditional probabilities, an alternative measure of the degree of deviation from serial independence, which can be used to compare across players, can be constructed based on the number of runs in a sequence. Let r_i^{dev} be the % deviation for each player in the number of runs in all point-games r_{pg_i} compared to the expected number of runs

FIGURE 1: Histogram of the individual player percentage deviation in the number of runs, r_i^{dev}



of a serially independent sequence $Exp\left(r_{pg_i}\right)$:

$$r_i^{dev} = E_{pg_i}\left(\frac{r_{pg_i} - Exp\left(r_{pg_i}\right)}{Exp\left(r_{pg_i}\right)}\right)$$

Figure 1 presents a histogram of the empirical distribution of r_i^{dev}. Over-alternation, i.e., switching too often or negative serial correlation, occurs if $r_i^{dev} > 0$ and under-alternation if $r_i^{dev} < 0$. Negative serial correlation is found for 69% of the players and the mean of r_i^{dev} is 5.55%, in line with the conclusions of over-alternation on average drawn from the empirical conditional probabilities. Furthermore, the 5th and 95th percentiles are large in magnitude, −14.6% and 23.7% respectively.

The power of the statistics conditional on this dataset is significantly higher than prior investigations, but varies according to the data available per player. Detailed simulations verifying the statistical power can be found in Appendix B. Based on these calculations, I refer to subjects with at least fifty matches as the high-power group, between twenty and fifty matches as the moderate-power group, and less than twenty matches as the low-power group. For the high-power group, the statistics have 80% power to detect an average effect size, and for the moderate-power group, 80% power to detect a higher (yet still plausible) effect size. Table 3 presents the statistics of all players who are represented in the high-power and moderate-power groups. The high power group consists of four players, three of whom were ranked Number 1 (Federer, Nadal, Djokovic) and the other Num-

ber 2 (Murray) in the world. The null hypothesis of serial independence is strongly rejected ($p < 0.005$) for all four players. The mean percentage deviation in the number of runs, r_i^{dev}, is −10.7%, 6.9%, −13.4%, and −3.4% respectively. Also, the probability of finding over-alternation in each player's point-game runs statistics is 0.25, 0.65, 0.2 and 0.42 respectively.

The moderate-power group of players consists of fourteen players, all of whom are Top 10 (career-high) ranked players with the exception of two players. The null hypothesis of serial independence is rejected for nine players. Notably, out of both power groups (high and moderate), four out of the five number 1 ranked players were found to exhibit serial dependence — the exception is Andre Agassi. These results are not sensitive to the grouping of players according to high- and moderate-power. Running the KS-tests on all players (including the low-power group) leads to a rejection of the null hypothesis of no serial correlation at the 5% level for 19.69% of the players. The B-H multiple-comparisons correction yields a rejection rate of 2.3% (nine players) — this correction is overly conservative due to the players in the low-power group. The rejections still include very highly ranked players (including No. 1), e.g., Federer, Nadal, and Djokovic. A table reporting all the player-level statistics can be found in the supplement.

Returning to the question of economic significance of the observed deviations, note that the mean number of runs per point-game is 16.5, i.e., roughly 33 per match. Therefore, the deviations from serial independence of the top-ranked players in the high-power group Federer, Nadal, Djokovic correspond to approximately 3.5 fewer runs, 2.3 more runs, and 4.4 fewer runs than expected respectively. The moderate-power group also includes players whose statistically significant deviations are approximately of the same magnitude — see Berdych, Ferrer and Dimitrov. These *individual* deviations may be difficult to detect for the average player, unless two players are matched up often enough, which is not unreasonable for the very best players. Furthermore, the *average population* tendency to over-alternate will be more readily detectable by attentive players due to the large number of observations. I return to this issue later in the manuscript.

4.2.1 Are within-subject deviations from serial (in)dependence a result of strategic best response to lower-ranked players or match characteristics?

Experimental and field studies have shown that beliefs about opponents' rationality can affect the likelihood of reaching the normative solution of a game. Palacios-Huerta & Volij (2009) find that the subgame-perfect equilibrium in the Centipede game is more likely to result if both players are expert chess players, less likely if chess players are matched with students, and least likely when students play other students. Similarly, Bosch-Domenech et al. (2002) find extensive iterated belief-based reasoning about the sophistication of opponents in a guessing game — many subjects who showed an understanding of the Nash equilibrium nevertheless chose to deviate. For example, suppose that the receiver is susceptible to the representativeness bias concerning random sequences or the law of small numbers (Tversky & Kahneman, 1971), i.e., believes that a switch in the serve direction is more likely after a sequence of the same serve directions. Consequently, even if the server is randomizing efficiently, the receiver will expect the sequence of the serves to over-alternate. The latter could be exploited by a server choosing to under-alternate, leading to an increase in the probability of mismatch in the sender and receiver directions, thereby increasing the probability of the server winning the point.

I examine whether such strategic deviations occur in tennis by using a cross-sectional regression model with fixed effects to absorb the between-subject variation leaving the within-subject variation to be modeled, i.e., within-player strategic adaptation to an opponent and/or match characteristics. The following independent variables are included. The current match rankings (not career-high) of both the server and the receiver (or opponent). The former captures within-player variation in randomization, which may occur as a result of the accumulation of experience/expertise (proxied by the player's own ranking). The latter represents the combined ability and expertise of the opponent. If lower-ranked players are more susceptible to the law of small numbers, then servers should deviate more from serial independence in the direction of under-alternating, the lower-ranked their opponent is. The current match rankings of both the server and receiver $(Rank_t^s, Rank_t^r)$ are transformed into $R_t^{own} = 8 - \log_2 Rank_t^s$ as suggested by Klaassen & Magnus (2009); the same transformation is used for R_t^{opp} where the subscript t denotes the current — not career high — ranking.

Other variables capture possibly important match characteristics. The length of a match, specifically the total number of points played, is included as the variable N_{points}. This variable could capture the effects of fatigue (and difficulty of the match) on the efficiency of serve randomization. The variable $\overline{L}_{rally}$ denotes the mean number of shots played per point, or the length of a rally. This variable could influence serve randomization in two possible ways. First, the greater the rally length, the more time that elapses between serves — consequently, a player who incorrectly conditions on prior behavior in a biased attempt to randomize, may actually benefit from greater rally lengths. Second, although the ranking of the opponent would capture the expected difficulty of a match, a greater length rally might indicate that this *specific* match differs in difficulty.[7] Consequently, this might increase the incentives for a player to exert more effort or greater care in randomizing efficiently. The vari-

[7]For example, the opponent ranking would not capture elements such as the effects of a recent injury, increased fatigue due to a busy schedule, the effects of different court surfaces et cetera.

TABLE 3: Individual player statistics

Name	Career-high ranking	Matches	Serves	Runs (K)	Runs (p)	r_i^{dev} (%)
High-power group						
Roger Federer	1	160	12,413	0.258	0.000	−10.7
Rafael Nadal	1	142	8,708	0.169	0.000	6.9
Novak Djokovic	1	126	8,984	0.345	0.000	−13.4
Andy Murray	2	68	4,935	0.154	0.003	−3.4
Moderate-power group						
Stanislas Wawrinka	3	42	3,035	0.180	0.007	−4.5
Tomas Berdych	4	38	2,244	0.370	0.000	13.9
David Ferrer	3	36	2,097	0.395	0.000	16.1
Milos Raonic	4	34	2,359	0.179	0.023	−7.5
Diego S. Schwartzman	57	31	2,078	0.196	0.014	4.1
Kei Nishikori	4	30	1,726	0.121	0.321	−3.6
Juan Martin DelPotro	4	29	2,025	0.174	0.054	2.4
Pete Sampras	1	28	2,260	0.272	0.000	−8.7
Bernard Tomic	17	26	1,961	0.136	0.266	−3.6
Richard Gasquet	7	24	1,535	0.150	0.211	3.2
Jo Wilfried Tsonga	5	23	1,480	0.209	0.030	−6.4
Grigor Dimitrov	8	23	1,484	0.482	0.000	18.4
Andre Agassi	1	22	1,693	0.138	0.343	2.0
Gilles Simon	6	21	1,524	0.298	0.001	9.4

able W_{diff} also captures the difficulty of a specific match as it is the difference between the rate at which the player won and lost points in a given match. If W_{diff} is close to zero, then the match is relatively even. Finally, the variable $\ln(Round)$ denotes the round of the match and is a proxy for the expected tournament payoff (factoring in the probability of winning) and the effects of stress or pressure in the later rounds. The variable *Round* is coded as follows: if the match is a final ($Round = 1$), semi-final ($Round = 2$), quarter-final ($Round = 3$), pre-quarter-final or Top 16 ($Round = 4$), or any lower qualifying round ($Round = 5$). Taking the logarithm of this scale imposes a concave relationship, i.e., that the effects of qualifying for each round have an increasingly larger additional effect through the increase in player incentives (monetary or otherwise).[8]

Note, however, that the causality of the variables $\left(N_{points}, \overline{L}_{rally}, W_{diff}\right)$ may also run in the opposite direction. That is, poor serve randomization could conceivably have direct effects on these variables. For example, if a server exhibits serial correlation and the receiver exploits this, then the receiver would be more likely to return the serve leading to longer rallies on average. Similarly, this could also affect the percentage of points won by the player or the number of points in the match. To remove the problem of endogeneity, I calculate $N_{points}, \overline{L}_{rally}, W_{diff}$ only using data where the player was the *receiver*, not the *server*.

The complete model is shown below in Equation 1, errors are normally distributed.[9] To allow for the possibility that players' strategic adaptation may depend on whether they are, on average, players who over- or under-alternate, the model estimates the set of coefficients separately for these two groups (the distinction is made on the basis of the sign of r_i^{dev}). The coefficients are denoted separately as β^+ for players where $r_i^{dev} > 0$ and β^- for players where $r_i^{dev} < 0$.

$$r_{pg_i}^{dev} = \alpha_i + \beta_1^{\pm} R_t^{own} + \beta_2^{\pm} R_t^{opp} + \beta_3^{\pm} N_{points} + \beta_4^{\pm} \overline{L}_{rally} + \beta_5^{\pm} W_{diff} + \beta_6^{\pm} \ln(Round) + \epsilon_{pg_i} \quad (1)$$

The results of the regression are displayed in Table 4 — the top half of the table presents the coefficients for players that

[8] Results are similar if instead the regression included dummy variables of the round, which however reduces the degrees of freedom.

[9] The conclusions are robust to the assumption of normally distributed errors, as bootstrapped standard errors did not alter the results. Also, because current match rankings were not available before 1980 in the database, this analysis is based on 4,336 point-games from 383 players; this is a minimal reduction compared to the total of 4,372 point-games and 391 players.

TABLE 4: The dependence of deviations in runs $r_{pg_i}^{dev}$ on player rankings and match characteristics

# of obs. 4336, # of players 383				
Independent var.: $F\,(12, 3941) = 0.38$, $p = 0.97$				
Fixed-effects: $F\,(382, 3941) = 1.19$, $p = 0.01$				
	Coeff.	s.e.	t	p
α	−0.165	1.998	−0.08	0.93
$r_i^{dev} > 0$				
$\beta_1^+ \left(R_t^{own}\right)$	0.057	0.436	0.13	0.90
$\beta_2^+ \left(R_t^{opp}\right)$	0.033	0.179	0.19	0.85
$\beta_3^+ \left(N_{points}\right)$	0.011	0.015	0.77	0.44
$\beta_4^+ \left(\overline{L}_{rally}\right)$	0.022	0.380	0.06	0.95
$\beta_5^+ \left(W_{diff}\right)$	−0.879	1.218	−0.72	0.47
$\beta_6^+ \left(\ln\left(Round\right)\right)$	0.728	0.722	1.01	0.31
$r_i^{dev} < 0$				
$\beta_1^- \left(R_t^{own}\right)$	−0.303	0.484	−0.63	0.53
$\beta_2^- \left(R_t^{opp}\right)$	0.054	0.213	0.26	0.80
$\beta_3^- \left(N_{points}\right)$	0.005	0.013	0.38	0.71
$\beta_4^- \left(\overline{L}_{rally}\right)$	0.405	0.438	0.92	0.36
$\beta_5^- \left(W_{diff}\right)$	−0.434	1.352	−0.32	0.75
$\beta_6^- \left(\ln\left(Round\right)\right)$	0.861	0.742	1.16	0.25

TABLE 5: Individual regressions of $r_{pg_i}^{dev}$ on player and match characteristics

Player	Federer	Nadal	Djokovic	Murray
# of obs.	320	284	252	136
F	0.66	2.15	1.57	0.14
p	0.68	0.048	0.037	0.99
R^2	0.012	0.045	0.037	0.006
Coeff. $(s.e.)$				
$\beta_1 \left(R_t^{own}\right)$	0.56	−2.80*	−0.19	0.61
	(0.93)	(1.23)	(1.60)	(2.44)
$\beta_2 \left(R_t^{opp}\right)$	0.43	0.97	−0.03	−0.50
	(0.57)	(0.74)	(0.78)	(0.88)
$\beta_3 \left(N_{points}\right)$	0.014	0.041	−0.062*	0.0004
	(0.026)	(0.033)	(0.028)	(0.043)
$\beta_4 \left(\overline{L}_{rally}\right)$	1.35	−0.59	−0.19	0.19
	(1.14)	(1.12)	(0.93)	(1.35)
$\beta_5 \left(W_{diff}\right)$	−0.40	−5.45	−9.45*	−0.28
	(3.79)	(4.83)	(4.49)	(5.17)
$\beta_6 \left(\ln\left(Round\right)\right)$	1.97	4.51	−0.52	0.46
	(1.77)	(2.38)	(2.19)	(3.22)
α	−24.48*	18.33	−5.86	−5.86
	(8.37)	(10.98)	(14.28)	(17.85)

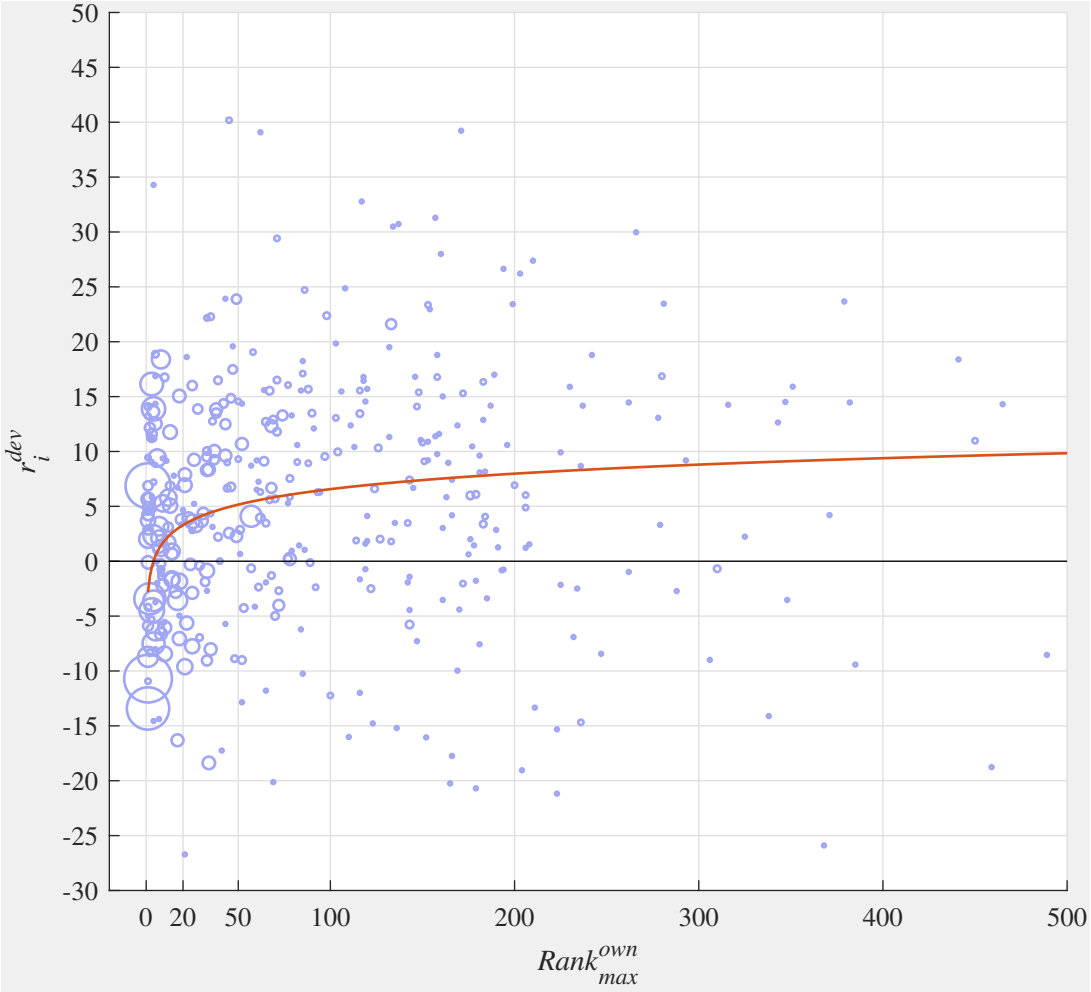
* denotes significance at the 5% level.

over-alternate on average, the bottom half those that under-alternate. A joint F-test of the null hypothesis that the set of independent variables are not different from zero cannot be rejected ($F\,(12, 3941) = 0.38, p = 0.97$). Similarly, tests for each independent variable cannot be rejected at the 5% level. I conclude that players do not systematically strategically manipulate their serve randomization according to the rank of their opponent, and furthermore that there is also no significant effect of match characteristics on behavior. This finding is robust to adding interactions between R_t^{own} and the other variables in the regression model, which would allow sensitivity to match characteristics to depend on a player's own rank — see Table 8 (Appendix D) for the regression results. Importantly, there is significant heterogeneity in r_i^{dev} between players as captured by the estimated fixed-effects ($F\,(382, 3941) = 1.19, p = 0.01$).

Table 5 presents individual regressions using the same set of regressors as above for the players in the high-power group. These regressions allow for the possibility of heterogeneity not only in the fixed-effects but also in the estimated coefficients of the independent variables. For example, it is possible that top-ranked players may adapt strategically to their opponent or the characteristics of each match, but lower ranked players may lack this ability. The prior regression on the whole set of players may thus have masked this heterogeneity. Note, that the bulk of the observations of R_t^{own} in these individual regressions fall within the Top 10 ranking range. Therefore, conclusions regarding the within-subject variation in randomization with rank are valid only within this range — it is possible that learning more efficient randomization may occur at much lower rankings. By contrast, there is significant variation in R_t^{opp} allowing for more general conclusions.

From Table 5, none of the variables are statistically significant for Federer and Murray; however, the $\beta_1 \left(R_t^{own}\right)$ coefficient for Nadal and $\beta_3 \left(N_{points}\right)$ and $\beta_5 \left(W_{diff}\right)$ coefficients for Djokovic are significantly different from zero ($p = 0.023, 0.025$ and 0.036, respectively). Of course, by increasing the sample size enough it is possible to reject any hypothesis for an arbitrarily small effect size. Therefore, the economic significance, or effect size, of the deviations is important — if they are small, then we should be cautious in concluding that players are not serving optimally even if statistical significance is found. Relatively small deviations may be either too difficult or too costly to detect and/or exploit. The economic significance, or effect size of these coefficients is more clearly illustrated by ω^2

FIGURE 2: Estimated weighted regression of the relationship between r_i^{dev} and career-high rank (circle size is proportional to the number of point-games)



or converting them to standardized beta coefficients.[10] For Nadal, ω^2 for β_1 (R_t^{own}) is equal to 0.015. For Djokovic, ω^2 for β_3 $\left(N_{points}\right)$ and β_5 $\left(W_{diff}\right)$ is 0.016 and 0.014, respectively. Consequently, I conclude that, while statistically significant, these within-subject findings explain very little variation, particularly compared to the between-subject (unconditional) deviations from serial independence for these players found above. In conjunction with the insignificant findings in the panel regression (Table 4), I conclude that there is no significant and systematic evidence of the existence of strategic deviations conditional either on the opponent's rank or the characteristics of a match.

4.2.2 Are between-subject deviations from serial independence dependent on a player's own career-high ranking?

Since the previous results have ruled out any systematic within-subject variation in serve randomization, in this section I focus solely on the between-player variation after averaging the $r_{pg_i}^{dev}$ observations into the player averages r_i^{dev}. Figure 2 shows the estimated function for all players relating $r_i^{dev} = \delta_0 + \delta_1 R_{max}^{own} + \epsilon_{pg_i}$, where $R_{max}^{own} = 8 - \log_2 Rank_{max}^{own}$, where the subscript max indicates the career-high rank.[11] The coefficient $\delta_0 + 8\delta_1$ corresponds to the mean value of r_i^{dev} for No. 1 ranked players. To account for the different number of observations determining r_i^{dev} for each player i, a weighted regression is employed with weights proportional to the number of point-games available for each player. Ro-

[10] For Nadal, the standardized coefficient for β_1 $\left(R_t^{own}\right)$ is equal to -0.143. For Djokovic, the beta coefficients for β_3 $\left(N_{points}\right)$ and β_5 $\left(W_{diff}\right)$ are -0.147 and -0.134, respectively.

[11] Using instead linear and power law funcions of the rank in this regression led to higher RMSE, confirming the suggestion to use this logarithmic transformation by Klaassen & Magnus (2009).

TABLE 6: Regressions of r_i^{dev} on the rank of players.

Regression	(1)	(2)	(3)	(4)
Player ranks included:	All	Top 100	Top 20	Top 10
Obs.	391	229	98	75
F(1,389)	61.24	33.46	7.75	9.51
p	0.000	0.000	0.007	0.003
RMSE	9.06	8.7	8.66	8.86
δ_1	−1.41	−1.51	−1.61	−2.64
s.e.	(0.18)	(0.26)	(0.58)	(0.86)
p	0.000	0.000	0.006	0.003
δ_0	8.49	9.12	9.89	17.54
s.e.	(0.98)	(1.49)	(3.82)	(6.01)
p	0.000	0.000	0.01	0.005
$\delta_0 + 8\delta_1$	−2.76	−2.96	−2.95	−3.59
s.e.	(0.74)	(0.91)	(1.25)	(1.38)
p	0.000	0.001	0.02	0.01

bustness tests were performed by running the same regression not only on the whole set of players, but also the Top 100, Top 20 and Top 10 players separately — see regressions (1)–(4) in Table 6. In all regressions, the estimate δ_1 was negative and statistically different from zero, i.e., players were increasingly less prone to under-alternating and more prone to over-alternating the lower ranked they were. Indicatively, the $E_i\left[r_i^{dev} \mid Rank_{max}^{own}\right]$ conditional on player rankings $[1, 10, 20, 50, 100, 500]$ are $[-2.8, 1.9, 3.3, 5.2, 6.6, 9.8]$ respectively (for the regression including all players) — see the plotted regression fit in Figure 2. In all the regressions, the value of $E_i\left[r_i^{dev} \mid Rank_{max}^{own} = 1\right]$ is negative and significantly different from zero.[12] GPW also find that more highly ranked men players' behavior is closer to equilibrium, but their estimated logit regression on serve directions does not imply under-alternation on average for No. 1 ranked players.

The group of players ranked No. 1 and No. 2 therefore exhibit an *average* tendency to under-alternate, although at the individual player level analysis above, we rejected serial independence for some Top 2 players both because of under- and over-alternating. Although under-alternating serves is not an equilibrium strategy, if the majority of players are over-alternating as receivers, then this would be consistent with a best-response to the population of receivers. Recall however, that no evidence was found of conditioning server randomization on the opponent's rank, so players would have to be learning deviations at the *population* level. Unfortunately, this cannot be directly tested because the receivers' actions are not easily observable. They would depend on the exact position of the player in the court (further to the left or right), also the grip they are using on the tennis racket, i.e., whether it is more appropriate for a backhand or forehand shot, and any other preparation to receive the serve whether mental or physical.

[12]This is robust to the exclusion of all other data, as the weighted mean of r_i^{dev} for No. 1 ranked players only, is equal to -4.1% with 95% confidence intervals −8.1% and 0.0%. Similarly, these statistics for Top 2 ranked players are −3.75% with 95% confidence intervals −7.15% and −0.35%; from Top 3 players onwards, the 95% confidence interval does not lie solely in the negative domain.

5 Conclusion

Using a new dataset with sufficient power to efficiently investigate the serial dependence in serve directions, I resolve the striking difference in the conclusions drawn by Walker & Wooders (2001) and Hsu et al. (2007) with respect to the serial (in)dependence of tennis serves. I corroborate the conclusion of the former study that there exist statistically significant deviations from serial independence in serves. Importantly, serial independence has been rejected even for players ranked Number 1 in the world at some point in their careers such as Federer, Nadal, and Djokovic. Over-alternation, or switching too often (negative serial correlation), was found to be more prevalent than under-alternation in the whole group of players — this is in line with the earlier results of the literature both in the laboratory and the field. Interestingly, Top 2 players were found to under-alternate on average — this would be a best response to a belief that the majority of tennis players tend to over-alternate in their direction as receivers. Furthermore, the lower the ranking of a player, the higher the degree of expected over-alternation. Within-player analyses did not find evidence of strategic deviations from serial independence by higher-ranked players when competing against lower-ranked players. Consequently, the observed serial dependence cannot be explained away as a rational response to non-equilibrium behavior of *individual* lower-ranked players with less experience and/or ability than the top players. These deviations might be difficult to detect and exploit at the level of each individual player, or within a single match, due to the small number of datapoints available for inference. However, learning the population-level tendency (outside of the Top 2 players) to over-alternate should be feasible and is one possible strategic explanation for the Top 2 players under-alternating on average. This is backed by extensive laboratory evidence that subjects playing repeated constant-sum games are capable of learning and exploiting the serial dependencies in their opponent's behavior given enough rounds of play (Spiliopoulos, 2012, 2013a,b, 2018). Future work could be directed at ascertaining whether the observed magnitude of deviations from randomness are easily detectable given the samples sizes observed in tennis matches and whether doing so would lead to an economically important advantage for a player. The latter would

require a formal model associating deviations from perfect randomization with the probability of winning points and ultimately the whole match. Also, more data covering the whole career span of players would allow for more powerful tests of within-player learning of randomization behavior. Finally, match characteristics proxying for the difficulty of a match, fatigue, induced pressure and incentives were not found to systematically influence the randomization behavior of players.

References

Bar-Hillel, M. & Wagenaar, W. A. (1991). The perception of randomness. *Advances in applied mathematics, 12*(4), 428–454.

Binmore, K. G., Swierzbinski, J., & Proulx, C. (2001). Does Minimax Work? An Experimental Study. *Economic Journal, 111*(473), 445–464.

Bloomfield, R. (1994). Learning a mixed strategy equilibrium in the laboratory. *Journal of Economic Behavior & Organization, 25*(3), 411–436.

Bosch-Domenech, A., Montalvo, J. G., Nagel, R., & Satorra, A. (2002). One, Two, (Three), Infinity, ...: Newspaper and Lab Beauty-Contest Experiments. *American Economic Review, 92*(5), 1687–1701.

Brown, J. N. & Rosenthal, R. W. (1990). Testing the Minimax Hypothesis: A Re-Examination of O' Neill's Game Experiment. *Econometrica, 58*(5), 1065–1081.

Budescu, D. V. (1987). A Markov model for generation of random binary sequences. *Journal of Experimental Psychology: Human Perception and Performance, 13*(1), 25–39.

Budescu, D. V. & Rapoport, A. (1994). Subjective randomization in one-and two-person games. *Journal of Behavioral Decision Making, 7*(4), 261–278.

Buzzacchi, L. & Pedrini, S. (2014). Does player specialization predict player actions? Evidence from penalty kicks at FIFA World Cup and UEFA Euro Cup. *Applied Economics, 46*(10), 1067–1080.

Chiappori, P. A., Levitt, S., & Groseclose, T. (2002). Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. *American Economic Review, 92*(4), 1138–1151.

Clegg, B. A., DiGirolamo, G. J., & Keele, S. W. (1998). Sequence learning. *Trends in Cognitive Sciences, 2*(8), 275–281.

Coloma, G. (2007). Penalty Kicks in Soccer An Alternative Methodology for Testing Mixed-Strategy Equilibria. *Journal of Sports Economics, 8*(5), 530–545.

Dohmen, T. & Sonnabend, H. (2016). Further Field Evidence for Minimax Play. *Journal of Sports Economics*, (pp. 1–18).

Emara, N., Owens, D. M., Smith, J., & Wilmer, L. (2014). Minimax on the Gridiron: Serial Correlation and Its Effects on Outcomes in the National Football League. http://ssrn.com/abstract=2401325.

Erev, I. & Roth, A. E. (1998). Predicting How People Play Games : Reinforcement Learning in Experimental Games with Unique , Mixed Strategy Equilibria. *American Economic Review, 88*(4), 848–881.

Falk, R. & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*(2), 301–318.

Farmer, G. D., Warren, P. A., & Hahn, U. (2017). Who "believes" in the Gambler's Fallacy and why? *Journal of Experimental Psychology: General, 146*(1), 63–76.

Gauriot, R., Page, L., & Wooders, J. (2016). Nash at Wimbledon: Evidence from Half a Million Serves. https://papers.ssrn.com/sol3/Delivery.cfm?abstractid=2801877.

Hahn, U. & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review, 116*(2), 454–461.

Hsu, S.-H., Huang, C.-Y., & Tang, C.-T. (2007). Minimax Play at Wimbledon: Comment. *American Economic Review, 97*(1), 517–523.

Ioannou, C. A. & Romero, J. (2014). A generalized approach to belief learning in repeated games. *Games and Economic Behavior*, 87, 178–203.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Kareev, Y. (1992). Not that bad after all: Generation of random sequences. *Journal of Experimental Psychology: Human Perception and Performance, 18*(4), 1189–1194.

Kareev, Y. (1995). Through a narrow window: working memory capacity and the detection of covariation. *Cognition, 56*(3), 263–269.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review, 107*(2), 397–402.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a Narrow Window: Sample Size and the Perception of Correlation. *Journal of Experimental Psychology: General, 126*(3), 278–287.

Klaassen, F. J. G. M. & Magnus, J. R. (2009). The efficiency of top agents: An analysis through service strategy in tennis. *Journal of Econometrics, 148*(1), 72–85.

Kovash, K. & Levitt, S. D. (2009). Professionals Do Not Play Minimax: Evidence from Major League Baseball and the National Football League. *NBER Working Paper*.

Levitt, S. D., List, J. A., & Reiley, D. H. (2010). What Happens in the Field Stays in the Field: Exploring Whether Professionals Play Minimax in Laboratory Experiments. *Econometrica, 78*(4), 1413–1434.

Lopes, L. L. & Oden, G. C. (1987). Distinguishing between random and nonrandom events. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 13(3), 392–400.

Magnus, J. R. & Klaassen, F. J. G. M. (1999). On the advantage of serving first in a tennis set: Four years at Wimbledon. *The Statistician, 48*(2), 247–256.

Marsaglia, G., Tsang, W. W., & Wang, J. (2003). Evaluating Kolmogorov's Distribution. *Journal of Statistical Software, 8*(1), 1–4.

McGarrity, J. P. & Linnen, B. (2010). Pass or Run: An Empirical Test of the Matching Pennies Game Using Data from the National Football League. *Southern Economic Journal, 76*(3), 791–810.

Neuringer, A. (1986). Can people behave "randomly?": The role of feedback. *Journal of Experimental Psychology: General, 115*(1), 62–75.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review, 109*(2), 330–357.

Nyarko, Y. & Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica, 70*(3), 971–1005.

Ochs, J. (1995). Games with Unique, Mixed Strategy Equilibria: An Experimental Study. *Games and Economic Behavior, 10*(1), 202–217.

Okano, Y. (2013). Minimax play by teams. *Games and Economic Behavior, 77*(1), 168–180.

O'Neill, B. (1987). Nonmetric test of the minimax theory of two-person zerosum games. *Proceedings of the National Academy of Sciences of the United States of America, 84*(7), 2106–2109.

Oskarsson, A. T., Van Boven, L., McClelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin, 135*(2), 262–285.

Palacios-Huerta, I. (2003). Professionals play minimax. *The Review of Economic Studies*, 70, 395–415.

Palacios-Huerta, I. & Volij, O. (2008). Experientia Docet: Professionals Play Minimax in Laboratory Experiments. *Econometrica, 76*(1), 71–115.

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

Plott, C. R. (1996). Rational Individual Behavior in Markets and Social Choice Processes: The Discovered Preference Hypothesis. In K. J. Arrow, E. Colombatto, M. Perlman, & C. Schmidt (Eds.), *The Rational Foundations of Economic Behaviour* (pp. 225–250). London: Palgrave Macmillan.

Rapoport, A. & Budescu, D. V. (1992). Generation of random series in two-person strictly competitive games. *Journal of Experimental Psychology: General, 121*(3), 352–363.

Rapoport, A. & Budescu, D. V. (1997). Randomization in individual choice behavior. *Psychological Review, 104*(3), 603–617.

Remillard, G. & Clark, J. M. (2001). Implicit learning of first-, second-, and third-order transition probabilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 483–498.

Roth, A. E. & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior, 8*(1), 164–212.

Shachat, J. M. (2002). Mixed Strategy Play and the Minimax Hypothesis. *Journal of Economic Theory, 104*(1), 189–226.

Shachat, J. M. & Swarthout, J. (2004). Do we detect and exploit mixed strategy play by opponents? *Mathematical Methods of Operational Research, 59*(3), 359–373.

Spiliopoulos, L. (2012). Pattern recognition and subjective belief learning in a repeated constant-sum game. *Games and Economic Behavior, 75*(2), 921–935.

Spiliopoulos, L. (2013a). Beyond fictitious play beliefs: Incorporating pattern recognition and similarity matching. *Games and Economic Behavior*, 81, 69–85.

Spiliopoulos, L. (2013b). Strategic adaptation of humans playing computer algorithms in a repeated constant-sum game. *Autonomous Agents and Multi-Agent Systems, 27*(1), 131–160.

Spiliopoulos, L. (2018). The determinants of response time in a repeated constant-sum game: A robust Bayesian hierarchical model. *Cognition*, 172, 107–123.

Sun, Y. & Wang, H. (2010). Perception of randomness: On the time of streaks. *Cognitive Psychology, 61*(4), 333–342.

Sun, Y. & Wang, H. (2011). Probability theory and perception of randomness: Bridging "ought" and "is". *Behavioral and Brain Sciences, 34*(05), 271–272.

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105–110.

Van Essen, M. & Wooders, J. (2015). Blind stealing: Experience and expertise in a mixed-strategy poker experiment. *Games and Economic Behavior, 91*(C), 186–206.

Walker, M. & Wooders, J. (2001). Minimax play at Wimbledon. *American Economic Review, 91*(5), 1521–1538.

Walker, M., Wooders, J., & Amir, R. (2011). Equilibrium play in matches: Binary Markov games. *Games and Economic Behavior, 71*(2), 487–502.

Warren, P. A., Gostoli, U., Farmer, G. D., El-Deredy, W., & Hahn, U. (2018). A re-examination of "bias" in human randomness perception. *Journal of Experimental Psychology: Human Perception and Performance, 44*(5), 663–680.

Wooders, J. (2010). Does Experience Teach? Professionals and Minimax Play in the Lab. *Econometrica, 78*(3), 1143–1154.

Zhao, J., Hahn, U., & Osherson, D. (2014). Perception and identification of random events. *Journal of Experimental Psychology: Human Perception and Performance, 40*(4), 1358–1371.

TABLE 7: Power and size calculations (values are symmetric around $q^{LL} = q^{RR} = 0.5$)

	$q^{LL} = q^{RR}$										
# of matches	0.40	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.50
1	0.345	0.296	0.243	0.197	0.157	0.122	0.098	0.075	0.060	0.050	0.052
10	0.991	0.973	0.931	0.853	0.729	0.584	0.401	0.251	0.139	0.068	0.050
20	1.000	1.000	0.999	0.989	0.948	0.860	0.689	0.435	0.222	0.093	0.049
50	1.000	1.000	1.000	1.000	1.000	0.998	0.970	0.815	0.476	0.155	0.052
100	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.986	0.775	0.272	0.048
150	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.917	0.375	0.050

Appendix A: Statistical tests

Serial dependence for each point-game in the data is performed using the two-sided exact runs test. Let n_L and n_R denote the number of serves to the left and right in a single point-game, respectively; let n_r denote the number of runs in the sequence of $n_L + n_R$ serves. The probability $q_{pg}(r)$ of finding r runs in such a sequence is given by Equation 2 below; note, $Q_{pg}(r)$ denotes the cumulative distribution.

$$q_{pg}(r) = \begin{cases} \dfrac{2\binom{n_L-1}{\frac{r}{2}-1}\binom{n_R-1}{\frac{r}{2}-1}}{\binom{n_L+n_R}{n_L}} & r \text{ is even} \\ \dfrac{\binom{n_L-1}{\frac{(r-1)}{2}}\binom{n_R-1}{\frac{(r-1)}{2}-1}}{\binom{n_L+n_R}{n_L}} + & \\ \dfrac{\binom{n_R-1}{\frac{(r-1)}{2}}\binom{n_L-1}{\frac{(r-1)}{2}-1}}{\binom{n_L+n_R}{n_L}} & r \text{ is odd} \end{cases} \tag{2}$$

Randomized test statistics are generated according to Walker & Wooders (2001, p. 1533), to satisfy the requirements of the Kolmogorov-Smirnov test, namely that they be identically and independently distributed, and possess a continuous cumulative distribution function. The randomized test statistic t_{pg} is a draw from the uniform distribution $U\left[Q_{pg}(r), Q_{pg}(r-1)\right]$.

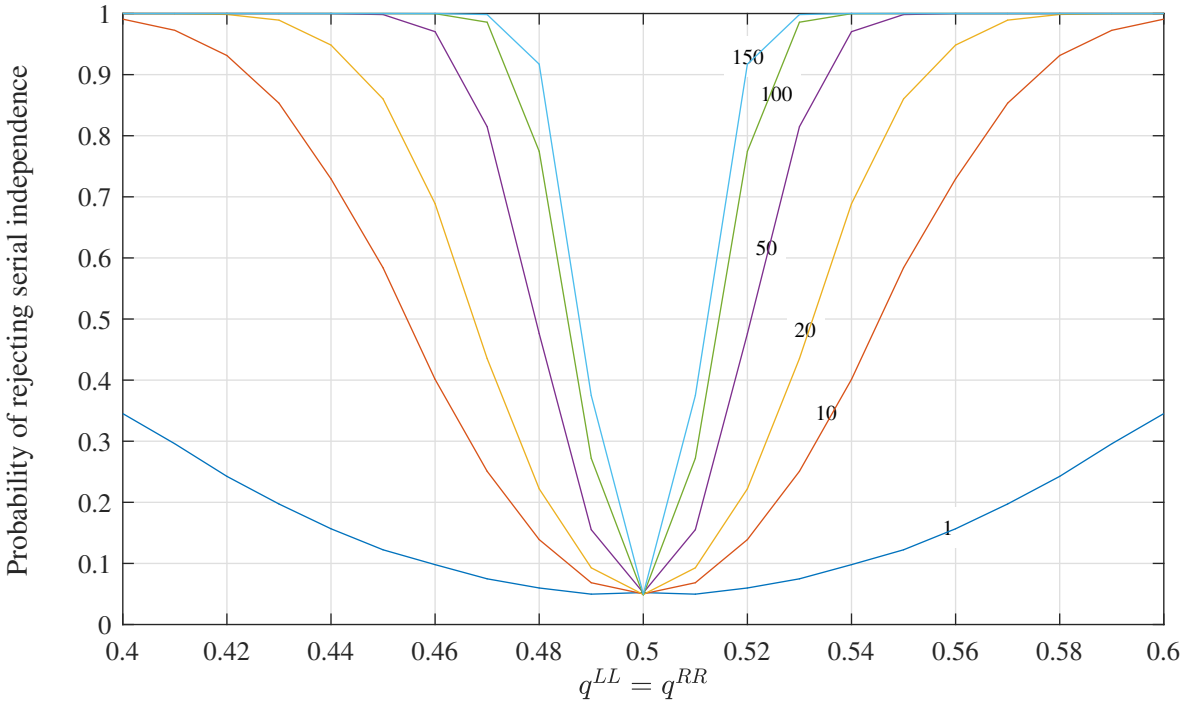
The individual point-game level test is a KS-test on the distribution of the randomized (exact run) test statistics at the point-game level for all the players — this is the test in WW and HHT. For the player-level analysis it is the Kolmogorov-Smirnov test on the distribution of point-game statistics pg_i for each player only. The latter permits the testing of serial independence for each player rather than the set of players. All Kolmogorov-Smirnov tests were implemented by the kstest function in Matlab, based on the algorithm in Marsaglia et al. (2003) capable of computing p-values with 13-15 digit accuracy for $n \geq 2$ samples. The reported critical values are $K = \max_{t_{pg}} \left|F\left(t_{pg}\right) - t_{pg}\right|$, where $F\left(t_{pg}\right)$ is the cumulative distribution function of the (randomized) exact runs test statistic t_{pg} (or t_{pg_i} for the player-level analysis) — note, the reported critical values are not scaled by $\sqrt{n}$, as is often done when using the asymptotic distribution.

Appendix B: Statistical power calculations

The amount of data for each player varies greatly in the dataset, from only one match to 160 matches. Consequently, the power of the statistical tests will also vary significantly by player. I present approximate power and size calculations of the proposed player-level test below. To simplify the presentation of the calculations, I perform these on a single "representative" point-game. From Table 2, the probability of serving left and right are in the vicinity of 0.5 across the ad and deuce point-games. Therefore, I perform the calculations assuming that $q^L = q^R = 0.5$, and varying the probabilities $q^{LL} = q^{RR}$ from 0.4 to 0.6, which includes the range of empirical estimates found in the data. Figure 3 displays the results from 10,000 simulated draws for players with varying numbers of matches ranging from 1 to 150 (the maximum in the dataset is 160 matches for Federer) and assuming 65 (left or right) serves per match (the average observed in the data). Table 7 presents the exact values.

The probability of rejecting the null hypothesis of serial independence (at the 5% significance level) is shown on the y-axis against different effect sizes on the x-axis (deviations from serial dependence as determined by $q^{LL} - q^L = q^{RR} - q^R$). If $q^{LL} = q^{RR} = 0.5$, then the simulated data exhibits serial independence. Therefore, the probability of (incorrectly) rejecting H_0 is equivalent to the size of the test, which should be 5%. As can be seen in the figure, the test has the appropriate size regardless of the amount of serve data available for a player. For all other values of $q^{LL} = q^{RR}$, the

FIGURE 3: Power and size calculations conditional on the number of matches



curves specify the power of the test, i.e., of correctly rejecting H_0. I compute an approximate expected effect size from the empirical data in the following way. Define the average effect size as the mean of the differences $\left|q^L - q^{LL}\right|$ and $\left|q^R - q^{RR}\right|$ for both point-games of the estimates presented in Table 2 — this is roughly 0.03.[13] Therefore, treating this as the expected effect size and under the assumption that $q^L = q^R = 0.5$, the corresponding conditional probabilities associated with the expected effect size are $q^{LL} = q^{RR} = 0.47$ and 0.53. The evolved norm in the literature for adequate power is 80%; this can be achieved at $q^{LL} = q^{RR} = 0.47$ with data from 50 matches. For a larger effect size at $q^{LL} = q^{RR} = 0.45$, data from 20 matches achieves a little more than 80% power. Based on these calculations, I refer to subjects with at least fifty matches as the high-power group, between twenty and fifty matches as the moderate-power group, and less than twenty matches as the low-power group. Note, that the findings in WW and HHT were based on twenty players from ten matches. Therefore, the approximate power of these studies to detect the expected effect size (at the population level) of 0.03 is 44%.[14] Therefore there is a significant chance that WW and HHT reached opposite conclusions due to the relatively lower statistical power of the tests.

Appendix C: Reconciling WW and HHT

The conflicting conclusions found in HHT and WW with respect to the serial (in)dependence of tennis serves deserve attention, especially since both studies used the same statistical tests. I conjecture that this difference may be attributed to sampling variation (in the selection of matches to include in a study) due to the relatively small number of matches used in these studies. Both WW and HHT had 40 point-games from ten men's matches. Suppose that only 40 point-games were randomly chosen from this study's dataset to perform the runs test. What is the probability of rejecting the null hypothesis using the same number of point-games as both WW and HHT? I resample 40 point-games for a total of 10,000 times from the complete set of point-games in the dataset (alternatively, only from point-games for players ranked No. 1), and calculate the probability of rejecting serial independence using these re-sampled datasets. Based on these calculations the probability of rejecting serial independence is 0.11 (or 0.36 for No. 1 ranked), i.e., in 11% (36%) of the

[13] For ad point-games, the differences are 0.537 – 0.51 = 0.027 and 0.463 – 0.43 = 0.33 for left and right serve directions respectively. For the deuce point-games, the differences are 0.486 – 0.461 = 0.025 and 0.514 – 0.487 = 0.027 respectively.

[14] In the power analysis, the number of matches refers to each player. In WW and HHT the analysis includes all twenty players from ten matches; therefore the appropriate curve in the figure is the one for twenty matches, not ten matches (which would be the case for an individual player analysis).

sub-samples. Consequently, the probability of two studies (of 40 point-games each) reaching the opposite conclusions, one rejecting and the other not rejecting the null is approximately 0.2 (or 0.46 for No.1). Note, that WW and HHT predominantly had very highly-ranked players including many No. 1 players; therefore, the value of 0.46 estimated from the No. 1 ranked players is likely the more accurate estimate. I conclude that the sampling variation hypothesis — lower power of the earlier datasets due to their small size — is a likely cause of the different results reached in WW and HHT. I note that the authors of both studies explicitly considered the power of their statistical tests for their datasets, but they were limited by the practical constraints of collecting and encoding the data from a large number of tennis matches, which is very time-consuming.

Appendix D: Other results

TABLE 8: The dependence of deviations in runs $r_{pg_i}^{dev}$ on player rankings and match characteristics (including interactions)

	Coeff.	s.e.	t	p
α	−0.265	3.859	−0.07	0.95
$r_i^{dev} > 0$				
R_t^{own}	−0.098	1.079	−0.09	0.93
R_t^{opp}	−0.168	0.374	−0.45	0.65
$R_t^{own} \times R_{opp}^t$	0.067	0.094	0.71	0.48
N_{points}	−0.030	0.031	−0.95	0.34
$R_t^{own} \times N_{points}$	0.009	0.006	1.50	0.14
$\overline{L}_{rally}$	1.127	0.785	1.44	0.15
$R_t^{own} \times \overline{L}_{rally}$	−0.273	0.166	−1.65	0.10
W_{diff}	0.428	2.441	0.18	0.86
$R_t^{own} \times W_{diff}$	−0.404	0.620	−0.65	0.52
$\ln(Round)$	0.077	1.455	0.05	0.96
$R_t^{own} \times \ln(Round)$	0.212	0.335	0.63	0.53
$r_i^{dev} < 0$				
R_t^{own}	−0.333	1.226	−0.27	0.79
R_t^{opp}	−0.515	0.499	−1.03	0.30
$R_t^{own} \times R_{opp}^t$	0.120	0.102	1.18	0.24
N_{points}	0.043	0.041	1.04	0.30
$R_t^{own} \times N_{points}$	−0.007	0.007	−1.02	0.31
$\overline{L}_{rally}$	−0.045	1.035	−0.04	0.97
$R_t^{own} \times \overline{L}_{rally}$	0.076	0.185	0.41	0.68
W_{diff}	1.261	3.235	0.39	0.70
$R_t^{own} \times W_{diff}$	−0.359	0.668	−0.54	0.59
$\ln(Round)$	2.004	1.807	1.11	0.27
$R_t^{own} \times \ln(Round)$	−0.188	0.337	−0.56	0.58