Expertise, Gender, and Equilibrium Play*

Romain Gauriot† Lionel Page‡ John Wooders§

May 2019

Abstract

Mixed strategy Nash equilibrium is the cornerstone of our understanding of strategic situations that require decision makers to be unpredictable. Using data from nearly half a million serves over 3000 tennis matches, and data on player rankings from the ATP and WTA, we examine whether the behavior of professional tennis players is consistent with equilibrium. We find that win rates conform remarkably closely to the theory for men, but conform somewhat less neatly for women. We show that the behavior in the field of more highly ranked (i.e., better) players conforms more closely to theory.

*The authors are grateful to Guillaume Fréchette, Kei Hirano, Jason Shachat, Mark Walker for useful comments. Wooders is grateful for financial support from the Australian Research Council's *Discovery Projects* funding scheme (project number DP140103566). Gauriot is grateful for financial support from the Australian Research Council's *Discovery Projects* funding scheme (project number DP150101307).

†Division of Social Science, New York University Abu Dhabi, United Arab Emirates, romain.gauriot@nyu.edu.

‡Economics Discipline Group, School of Business, University of Technology Sydney, lionel.page@uts.edu.au.

§Division of Social Science, New York University Abu Dhabi, United Arab Emirates, john.wooders@nyu.edu.

1 Introduction

Laboratory experiments have been enormously successful in providing tightly controlled tests of game theory. The results of these experiments, however, have not been supportive of the theory for games with a mixed-strategy Nash equilibrium: student subjects do not mix in the equilibrium proportions and subjects exhibit serial correlation in their choices rather than the serial independence predicted by the theory. While the rules of an experimental game which requires players to be unpredictable may be simple to understand, it is far more difficult to understand how to play *well*. Student subjects no doubt understand the rules, but they have neither the experience, the time, nor the incentive to learn to play well. In professional sports, by contrast, players have typically devoted their lives to the game and they have substantial financial incentives, and thus it provides an ideal setting to test theory.

The present paper examines whether the behavior of sports professionals conforms to theory by combining a unique dataset from Hawk-Eye, a computerized ball tracking system employed at Wimbledon and other top championship tennis matches, with data on player rankings from the ATP (Association of Tennis Professionals) and the WTA (Women's Tennis Association). It makes several contributions: With a large dataset and a new statistical test we introduce, it provides a far more powerful test of the theory than in any prior study. It also provides a broad test of the theory by analysing the play of both men and women players with different degrees of expertise. It finds substantial differences in the degree to which the behavior of men and women conform to equilibrium. Most significantly, it shows that even tennis professionals differ in the degree to which their behavior conforms to theory and, remarkably, the on-court behavior of more highly ranked players conforms more closely to theory. We are aware of no similar result in the literature.

A critique of the results of prior studies using data from professional sports has been that they have low power to reject the theory.[1] Walker and Wooders (2001), henceforth WW, studies a dataset comprised of approximately 3000 serves made in 10 men's championship tennis matches. Chiappori, Levitt, and Groseclose (2002), henceforth CLG, and Palacios-Huerta (2003), henceforth PH, study 459 and 1417 penalty kicks, respectively. Our dataset, by contrast, contains the precise trajectory and bounce points of the tennis ball for nearly 500,000 serves from over 3000 profes-

[1] See Kovash and Levitt (2009).

sional tennis matches, and thereby provides an extremely powerful test of the theory. Camerer (2003) suggests that WW's focus on long matches, with the goal of generating a test with high statistical power, could introduce a selection bias in favor of equilibrium play. Our analysis does not suffer from this critique as it uses data from all the matches where the Hawk-Eye system was employed.

The large number of matches in our dataset requires the development of a novel statistical test for our analysis. When the number of points played in each match is small relative to the overall number of matches, as it is in our dataset, we show that a key statistical test employed in WW is not valid: even when the null hypothesis is true, the test rejects the null (implied Nash equilibrium) that winning probabilities are equalized across directions of serve. By contrast, the test that we develop, based on the Fisher exact test, rejects the true null hypothesis with exactly probability α at the α significance level. We show via Monte Carlo simulations that our test, as an added bonus, is substantially more powerful than the test used in WW and the subsequent literature.[2]

An unusual feature of our test is that the test statistic itself is random, and thus a different p-value is realized each time the test is conducted. It would be perfectly legitimate to run the test once and reject the null hypothesis if the p-value is less than the desired significance level. It is more informative, however, to report the empirical density of p-values obtained after running the test many times, and this is what we do. When reporting our results we will make statements such as "the empirical density of p-values places an $x\%$ probability weight on p-values below .05." Reporting the empirical density reveals the sensitivity of our conclusions to the randomness inherent in the test statistic. Since randomized tests are seldom used, we complement our analysis with the implementation of a deterministic test in Appendix B.[3] The deterministic test has low power in comparison to our randomized test.

We find that the win rates of male professional tennis players are strikingly consistent with the equilibrium play. Despite the enormous power of our statistical test – due to the large sample size and the greater power of the test itself – we can not reject the null hypothesis that winning probabilities are equalized across the direction of serve. We do not reject the null for either first or second serves. For first serves, the

[2]The WW test *is* valid for the data set it considered, where the number of points in each match was large relative to the number of matches, as we show in Section 6.

[3]We are grateful to Kei Hirano for suggesting the construction of the deterministic test.

empirical density of p-values places no probability weight on p-values below .05 (i.e., the joint null hypothesis is never rejected at the 5% significance level). For second serves, it places almost no probability weight on p-values below .05.

The win rates for female players, by contrast, conform somewhat less neatly to theory. The empirical density function of p-values places a 44.73% weight on p-values below .05 for first serves, and a 16.1% weight on p-values below .05 for second serves. Nonetheless, the behavior of female professional tennis players over 150,000 tennis serves conforms far more closely to theory than the behavior of student subjects in comparable laboratory tests of mixed-strategy equilibrium. Applying our test to the data from O'Neill's (1987) classic experiment, for example, we obtain an empirical density function of p-values that places probability one on p-values less than .05. Hence the null hypothesis that winning probabilities are equalized is resoundingly rejected based on the 5250 decisions of O'Neill's subjects while we obtain no such result for female professional tennis players, despite having vastly more data.

This result naturally raises the question of whether the behavior of better – more highly ranked – female tennis players conforms more closely to theory. To investigate the effect of ability on behavior, we divide our data into two subsamples based on the rank of the player receiving the serve. (It is important to keep in mind that it is the receiver's play that determines whether winning probabilities are equalized across directions of serve.) In one subsample the receiver is a "top" player, i.e., above the median rank, and in the other the receiver is a "non-top" player. We test the hypothesis that winning probabilities are equalized across the direction of serve on each subsample separately. For men, win rates conform closely to equilibrium on each subsample. This result is not surprising given the stunning conformity of behavior to theory in the overall sample for men.

As just noted, win rates conform to equilibrium somewhat less neatly for women. Significantly, in women's matches in which the receiver is a "top" player, we do not come close to rejecting the hypothesis that winning probabilities conform to equilibrium, while the equilibrium is resoundingly rejected for the subsample in which the receiver is not a top player. This result show that behavior of female receivers conforms more closely to theory for more highly ranked players.

What might explain this difference between male and female players? We conjecture that men's greater physical strength and corresponding faster serve causes men's payoffs in the contest for each point to be more sensitive to departures from

equilibrium play than in women's tennis. In our dataset, the average speed of the first serve for men is 160 kph, while for women it is 135 kph.[4] A receiver in men's tennis who fails to play equilibrium (and equalize the server's winning probabilities) is far more vulnerable to being exploited by the server. Consequently, there is a stronger selection pressure against male receivers who fail to equalize the server's winning probabilities than there is against female receivers.

A second implication of equilibrium theory is that the players' choices of direction of serve are random (i.e., serially independent) and hence unpredictable. We find that both male and female players exhibit serial correlation in their serves with female players' serves being significantly more serially correlated than male players' serves. We conjecture the difference is again result of the greater importance of the serve in men's tennis. In men's tennis, 8.71% of all first serves are "aces", with the receiver unable to place his racket on the ball. A male player whose serve is predictable surrenders a portion of the significant advantage that comes from having the serve. In women's tennis, by contrast, only 4.41% of first serves are aces. There is a stronger selection effect in favor of equilibrium in men's tennis.

Here too we find evidence that the behavior of higher ranked players conforms more closely to equilibrium: Higher-ranked male players exhibit less serial correlation in the direction of serve than lower ranked players. For female players, by contrast, rank does not have a statistically significant effect on the degree of serial correlation.

Related Literature

Our paper contributes to the literature investigating the degree to which the behavior of professions conforms to equilibrium. WW was the first paper to use data from professional sports to test the minimax hypothesis. It found that the win rates of male professional tennis players conformed to theory, in striking contrast to the consistent failure of subjects to follow the equilibrium mixtures (and equalize payoffs) in laboratory experiments. Even tennis players, however, exhibit negative serial correction in their choices, switching the direction of the serve too often to be consistent with random play, as predicted by the theory. von Neuman's notion of Minimax,

[4]For first serves by men, on average only 0.45 seconds elapses between the serve and the first bounce. The greater importance of the serve in men's tennis is evident from the fact that in men's tennis, the server wins 64% of all the points when he has the serve, while in women's tennis the server only wins 58% of the points.

the foundation of modern game theory, and Nash equilibrium coincide in two-player constant sum games, and we will use minimax and equilibrium interchangeably.

Hsu, Huang, and Tang (2007), henceforth HHT, broaden the analysis of WW. They find that win rates conformed to the theory for a sample of 9 women's matches, 8 junior's matches, and 10 men's matches. The greater power of our statistical test means that it potentially overturns their conclusions and indeed in some instances it does. Our test, applied to their data for women and juniors, puts weights of 18.1% and 49.2%, respectively, on p-values of less than .05. On the other hand, applying our test to WW's data or HHT's data for men, we reaffirm their findings that the behavior of male professional tennis players conforms to equilibrium. In both cases, the empirical density of p-values assigns zero probability to p-values below .05.

CLG study a dataset of every penalty kick occurring in French and Italian elite soccer leagues over a three year period (459 penalty kicks), and test whether play conforms to the mixed strategy Nash equilibrium of a parametric model of a penalty kick in which the kicker and goalkeeper simultaneously choose Left, Center, or Right. A challenge in using penalty kicks to test theory is that most kickers take few penalty kicks and, furthermore, a given kicker only rarely encounters the same goalie. The later is important since the contest between a kicker and goalie varies with the players involved, as do the equilibrium mixtures and payoffs.[5] CLG finds that the data conforms to the qualitative predictions of the model, e.g., kickers choose "center" more frequently than goalies.[6] A key contribution of CLG is the precise identification of the predictions of the equilibrium theory that are robust to aggregation across heterogeneous contests.

PH studies a group of 22 kickers and 20 goalkeepers who have participated in at least 30 penalty kicks over a five year period in a dataset comprised of 1417 penalty kicks. The null hypothesis that the probability of scoring is the same for kicks to the left and to the right is rejected at the 5% level for only 2 of the kickers.[7] Importantly, his analysis ignores that a kicker generally faces different goalkeepers (and different

[5] CLG provide evidence that payoffs in the 3×3 penalty kick game vary with the kicker, but not with the goalie.

[6] In a linear probability regression they find weak evidence against the hypothesis that kickers equalize payoffs across directions based on the subsample of 27 kickers with 5 or more kicks. This null is rejected at the 10% level for 5 of kickers, whereas only 2.7 rejections are expected.

[7] PH aggregates kicks to the center and kicks to a player's "natural side" and thereby makes the game a 2×2 game.

goalkeepers face different kickers) at each penalty kick.

In professional tennis, unlike soccer, we observe a large number of serves, taken in an identical situations (e.g., Federer serving to Nadal from the "ad" court), over a period of several hours.[8] The relationship between the players' actions and the probability of winning the point is the same in every such instance, and thus the data from a single match can be used to test equilibrium theory. There is no need to aggregate data as in CLG or PH.

The present paper is related to a literature that examines the effect of experience in the field on behavior in the laboratory (see, e.g., Cooper, Kagel, Lo and Gu (1999) and Van Essen and Wooders (2015)). Palacios-Huerta and Volij (2008) report evidence that professional soccer players behave according to equilibrium when playing abstract normal form games in the laboratory. Levitt, List, and Reiley (2011) are, however, unable to replicate this result, while Wooders (2010) argues that Palacios-Huerta and Volij (2008)'s own data is inconsistent with equilibrium. Levitt, List, and Sadoff (2011) show that expert chess players, who might expected to be skilled at backward induction reasoning, play the centipede game much like typical student subjects. Our work differs by examining the effect of expertise on the conformity of behavior in the field to equilibrium play.

Several papers have used data from professional sports to study the effect of pressure on behavior. Paserman (2010) finds evidence that player performance in professional tennis is degraded for more "important" points, i.e., points where winning or losing the point has a large influence on the probability of winning the match. Gonzalez-Diaz, Gossner, and Rogers (2012) find that players are heterogeneous in their response to important points and they develop a measure of a skill they call "critical ability." The probability of winning a point is highly responsive to the point's importance for players with high critical ability. Kocher, Lenz, and Sutter (2012) find for soccer that there is no first-mover advantage in penalty kick shootouts.

In Section 2 we present the model of a serve in tennis and the testable hypotheses implied by the theory. In Section 3 we describe our data. In Section 4 we describe our new statistical test of the hypothesis that winning probabilities are equalized, we present our results, and we show that the behavior of higher ranked players conforms more closely to theory than for lower ranked players. In Section 5 we report the results

[8]Typical experimental studies of mixed-strategy play likewise feature a fixed pair of players playing the same stage game repeatedly over a period of an hour or two.

of our test that the direction of serve is serially independent. In Section 6 we study the power of the WW test and our new test of the hypothesis that winning probabilities are equalized. We show that (i) the WW test is valid when the number of points in each match is large relative to the number of matches, but is not valid conversely, (ii) our new test is valid whether the number of matches is small (as in WW) or large, (iii) our new test is more powerful than the test used WW and subsequent studies, and we (iv) apply our test to the data from HHT. The appendices are intended for online publication.

2 Modelling The Serve in Tennis

We model each point in a tennis match as a 2×2 normal-form game. The server chooses whether to serve to the receiver's left (L) or the receiver's right (R). The receiver simultaneously chooses whether to overplay left or right. The probability that the server ultimately wins the point when he serves in direction s and the receiver overplays direction r is denoted by π_{sr}. Hence the game for a point is represented by Figure 1.

		Receiver L	Receiver R
Server	L	π_{LL}	π_{LR}
	R	π_{RL}	π_{RR}

Figure 1: The Game for a Point

Since one player or the other wins the point, the probability that the receiver wins the point is $1 - \pi_{sr}$, and hence the game is completely determined by the server's winning probabilities.

The probability payoffs in Figure 1 will depend on the abilities of the two players in the match and, in particular, on which player is serving. In tennis, the player with the serve alternates between serving from the ad court (the left side of the court) and from the deuce court (the right side). Since the players' abilities may differ when serving or receiving from one court or the other, the probability payoffs in Figure 1 may also depend upon whether the serve is from the ad or deuce court. At the first serve, the probability payoffs include the possibility that the server ultimately wins

the point after an additional (second) serve. Since the second serve is the final serve, the probability payoffs for a second serve will be different than those for a first serve.[9]

In addition to varying the direction of the serve, the server can also vary its type (flat, slice, kick, topspin) and speed. In a mixed-strategy Nash equilibrium, all types of serves which are delivered with positive probability have the same payoff. Therefore it is legitimate to pool, as we do, all serves of different types but in the same direction. Our test of the hypothesis that the probability of winning the point is the same for serves left and serves right can be viewed as a test of the hypothesis that all serves in the support of the server's mixture have the same winning probability.

We assume that within a given match the probability payoffs are completely determined by which player has the serve, whether the serve is from the ad or deuce court, and whether the serve is a first or second serve. Thus, there are eight distinct "point" games in a match. We assume that in every point game $\pi_{LL} < \pi_{LR}$ and $\pi_{RR} < \pi_{RL}$, i.e., the server wins the point with lower probability (and the receiver with higher probability) when the receiver correctly anticipates the direction of the serve. Under this assumption there is a unique Nash equilibrium and it is in (strictly) mixed strategies.

A tennis match is a complicated extensive form game: The first player to win at least four points and to have won two more points than his rival wins a unit of scoring called a "game." The first player to win at least six games and to have won two games more than his rival wins a "set." In a five-set match, the first player to win three sets wins the match. The players, however, are interested in winning points only in so far as they are the means by which they win the match. The link between the point games and the overall match is provided in Walker, Wooders, and Amir (2011) which defines and analyzes a class of games (which includes tennis) called Binary Markov games. They show that Nash equilibrium (and minimax) play in the match consists of playing, at each point, the equilibrium of the point game in which the payoffs are the winning probabilities π_{sr} of Figure 1. Thus play depends only on which player is serving, whether the point is an ad-court or a deuce-court point, and whether the serve is a first or second serve; it does not otherwise depend on the current score or

[9]If the first serve is a fault, then the server gets a second, and final, serve. If the second serve is also a fault, then server loses the point. First and second serves are played differently. In our data set, the average speed of a first serve for men is 160 kph and of a second serve is 126 kph (35.3% of first serves fault, but only 7.5% of second serves fault).

any other aspect of the history of play prior to that point.

Testing the Theory

Two testable implications come from the theory. The first is that a player obtains the same payoff from all actions which in equilibrium are played with positive probability. To illustrate this hypothesis, consider the hypothetical point game below.

		Receiver L	Receiver R	
Server	L	0.58	0.79	8/15
	R	0.73	0.49	7/15
		2/3	1/3	

Figure 2: An Illustrative Point Game

The receiver's equilibrium mixture is to choose L with probability 2/3 and R with probability 1/3. When the receiver follows this mixture, then the server wins the point with the same probability (*viz.* .65) whether he serves L or he serves R.[10] In an actual tennis match we do not observe the probability payoffs, and hence we can not compute the receiver's equilibrium mixture. Nonetheless, so long as the receiver plays equilibrium, then the server's probability of winning the point will be the same for serves L and serves R.

When theory performs poorly, an important question is whether the behavior of better players conforms more closely to theory. Does a better player, when receiving the ball, more closely follow his equilibrium mixture and therefore more closely equalize the server's winning probabilities? In Section 4 we provide evidence that top female players do equalize the servers' winning probabilities when receiving the ball, while lower ranked female players do not. Thus the behavior of better female players conforms more closely to theory. Both top and non-top male players equalize the server's winning probabilities when receiving.

The second implication of the theory, which comes from the equilibrium analysis of the extensive form game representing a match, is that the sequence of directions of the serve chosen by a server is serially independent. A server whose choices are serially

[10]The servers' equilibrium mixture also equalizes the receiver's winning probability for each of his actions. Since the receiver's action is not observed, we can not test this hypothesis.

correlated may be exploited by the receiver and therefore his play is suboptimal. Here we focus on the rank of the player serving the ball. Section 4 provides evidence that higher ranked male players exhibit less serial correlation in the direction of their serve than lower ranked males, and thus their behavior conforms more closely to theory.

3 The Data

Hawk-Eye is a computerized ball tracking system used in professional tennis and other sports to precisely record the trajectory of the ball. Our dataset consists of the official Hawk-Eye data for all matches played at the international professional level, where this technology was used, between March 2005 and March 2009.[11] Most of the matches are from Grand Slam and ATP (Association of Tennis Players) tournaments Overall, the dataset contains 3,172 different singles matches. Table 1 provides a breakdown of the match characteristics of our data.

		Female	Male	All
Surface	Carpet	35	174	209
	Clay	130	366	496
	Grass	95	204	299
	Hard	917	1251	2168
Best of	3	1177	1400	2577
	5	0	595	595
Events	Davis Cup (Fed Cup)	8	18	26
	Grand Slam	458	526	984
	Olympics	19	16	35
	ATP (Premier)	662	101	763
	International	-	473	473
	Master	-	825	825
	Hopman Cup	30	36	66
Total		1117	1995	3172

Table 1: Match Characteristics

[11] Hawk-Eye has been used to resolve challenges to line calls since 2006, which is evidence of the greater reliability of Hawk-Eye to human referees.

As the use of the Hawk-Eye system is usually limited to the main tournaments, the dataset contains a large proportion of matches from top tournaments (e.g., Grand Slams). Within tournaments, the matches in our dataset are more likely to feature top players as the Hawk-Eye system is used on the main courts and was often absent from minor courts at the time of our sample. As a consequence, the matches contained in the dataset tend to feature the best male and female players.

For each point played, our dataset records the trajectory of the ball, as well as the player serving, the current score, and the winner of the point.[12] When the server faults as a result of the ball failing to clear the net, then we extrapolate the path of the serve to identify where the ball would have bounced had the net not intervened. Figure 3 is a representation of a tennis court and shows the actual (in blue) and imputed (in red) ball bounces of first serves by men, for serves delivered from the deuce court. The dashed lines in the figure are imaginary lines – not present on an actual court – that divide the two "right service" courts and are used to distinguish left serves from right serves.

Our analysis focuses on the location of the first bounce following a serve. As is evident from Figure 3, serves are typically delivered to the extreme left or the extreme right of the deuce court. We classify the direction of a serve – left or right – from the server's perspective: A player serving from the left hand side of the court delivers a serve across the net into the receiver's right service court. A bounce above the dashed line (on the right hand side of Figure 3) is classified as a serve to the left, and a bounce below the dashed line is classified as a serve to the right. Likewise, for a player serving from the right hand side of the court, a bounce below the dashed line (on the left hand side of Figure 3) is classified as a left serve, while a bounce above is classified as a right serve.

One could more finely distinguish serve directions, e.g., left, center, and right, but doing so would not impact our hypothesis tests. So long as left and right are both in the support of the server's equilibrium mixture, serves in each direction have the

[12] Hawkeye records the path of the ball as a sequence of arcs between impacts of the ball with a racket, the ground, or the net. Each arc (in three dimensions) is decomposed into three arcs, one for each dimension – the x-axis, the y-axis, and the z-axis. Each of these arcs is encoded as a polynomial equation with time as a variable. For each arc in three dimensions we have therefore three polynomial equations (typically of degree 2 or 3) describing the motion of the ball in time and space.

same theoretical winning probability.

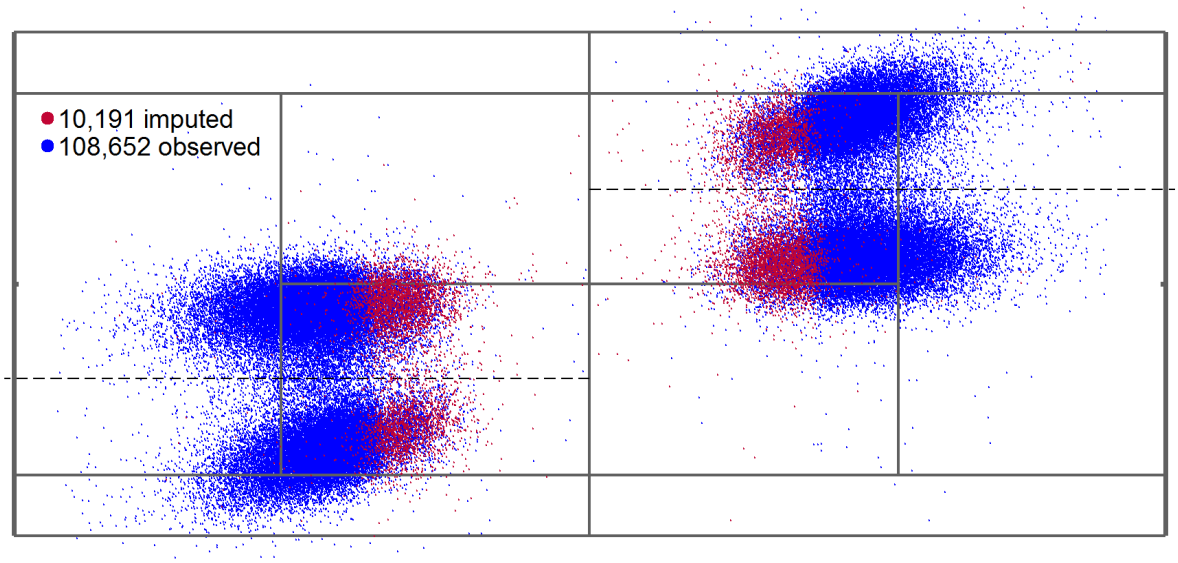


Figure 3: Ball Bounces for Deuce Court First Serves by Men

Second serves are delivered at slower speeds than first serves and are less likely to be a fault, but are also typically delivered to the left or right. See Appendix C, Figure C5.

While Hawk-Eye automatically records bounce data, the names of the players, the identity of the server and the score are entered manually. This leads to some discrepancies as a result of data entry errors. To ensure that the information we use in our analysis is correct, we check that the score evolved logically within a game: the game should start at 0-0, and the score should be 1-0 if the server wins the first point and 0-1 if the receiver wins the point. We do this for every point within a game. If there is even one error within a game, we drop the whole game. While conservative, this approach ensures that our results are based on highly accurate data. Table 2 reports the number of first and second serves for men and women that remain. A detailed description of the data cleaning process is provided in Appendix

A. We observe a total of 465,262 serves in the cleaned data.

Serve	Gender	Serves	Point Games
1st serve	Male	226,298	7,198
	Female	110,886	4,108
2nd serve	Male	86,702	7,198
	Female	41,376	4,108

Table 2: Number of Serves and Point Games

4 Testing for Equality of Winning Probabilities

According to equilibrium theory, the probability that the server wins the point is the same for serves left and for serves right.

Individual Play and The Fisher Exact Test

Let p_j^i denote the true, but unknown, probability that the server in point game i wins the point when the first serve is in direction j. We use the Fisher exact test to test the null hypothesis that $p_L^i = p_R^i = p^i$ for point game i, i.e., the probability that the server wins the point is the same whether serving to the left or to the right. Let $f(n_{LS}^i | n_S^i, n_L^i, n_R^i)$ denote the probability, under the null, that the server wins n_{LS}^i serves to the left, conditional on winning n_S^i serves in total, after delivering n_L^i and n_R^i serves to the left and to the right. As shown by Fisher (1935), this probability does not depend on p^i and is given by

$$f(n_{LS}^i | n_S^i, n_L^i, n_R^i) = \frac{\binom{n_L^i}{n_{LS}^i}\binom{n_R^i}{n_{RS}^i}}{\binom{n_L^i + n_R^i}{n_S^i}},$$

where $n_{RS}^i = n_S^i - n_{LS}^i$. Let $F(n_{LS}^i | n_S^i, n_L^i, n_R^i)$ be the associated *c.d.f.*, i.e.,

$$F(n_{LS}^i | n_S^i, n_L^i, n_R^i) = \sum\nolimits_{k=\max(n_S^i - n_R^i, 0)}^{n_{LS}^i} f(k | n_S^i, n_L^i, n_R^i).$$

In its standard application, the Fisher exact test rejects the null hypothesis at significance level α for n_{LS}^i such that $F(n_{LS}^i | n_S^i, n_L^i, n_R^i) \leq \alpha/2$ or $1 - F(n_{LS}^i - 1 | n_S^i, n_L^i, n_R^i) \leq \alpha/2$.

The Fisher exact test is the uniformly most powerful (UMP) unbiased test of the hypothesis that $p_L^i = p_R^i$ but, because of the discreteness of the density f, randomization is required to achieve a significance level of exactly α (see Lehmann and Romano (2005), p. 127). Let $\bar{n}_{LS}^i$ be the largest integer such that $F(\bar{n}_{LS}^i|n_S^i, n_L^i, n_R^i) \leq \alpha/2$ and $\underline{n}_{LS}^i$ be the smallest integer such that $1 - F(\underline{n}_{LS}^i - 1|n_S^i, n_L^i, n_R^i) \leq \alpha/2$. Without randomization, the true size of the test is only $F(\bar{n}_{LS}^i|n_S^i, n_L^i, n_R^i) + 1 - F(\underline{n}_{LS}^i - 1|n_S^i, n_L^i, n_R^i)$, which may be considerably smaller than α.

We implement a randomized Fisher exact test of exactly size α as follows: For each point game i, let t^i be the random test statistic given by a draw from the uniform distribution $U[0, F(n_{LS}^i|n_S^i, n_L^i, n_R^i)]$ if n_{LS}^i takes its minimum value, i.e., $n_{LS}^i = n_S^i - n_R^i$, and by a draw from the distribution $U[F(n_{LS}^i - 1|n_S^i, n_L^i, n_R^i), F(n_{LS}^i|n_S^i, n_L^i, n_R^i)]$ otherwise. Under the null hypothesis that $p_L^i = p_R^i$, the test statistic t^i is distributed $U[0, 1]$.[13] Hence rejecting the null hypothesis if $t^i \leq .025$ or $t^i \geq .975$ yields a test of exactly size .05.

The Fisher exact test and the randomized Fisher exact test make the same (deterministic) decision for all realizations of n_{LS}^i except $n_{LS}^i = \bar{n}_{LS}^i + 1$ or $n_{LS}^i = \underline{n}_{LS}^i - 1$.[14] If $n_{LS}^i = \bar{n}_{LS}^i + 1$, then the Fisher exact test does not reject the null, while the randomized test rejects it with probability

$$\frac{\alpha/2 - F(\bar{n}_{LS}^i|n_S^i, n_L^i, n_R^i)}{F(\bar{n}_{LS}^i + 1|n_S^i, n_L^i, n_R^i) - F(\bar{n}_{LS}^i|n_S^i, n_L^i, n_R^i)}.$$

Likewise, if $n_{LS}^i = \bar{n}_{LS}^i - 1$, the Fisher exact test does not reject the null, while the randomized test rejects it with probability

$$\frac{F(\bar{n}_{LS}^i - 1|n_S^i, n_L^i, n_R^i) - (1 - \alpha/2)}{F(\underline{n}_{LS}^i - 1|n_S^i, n_L^i, n_R^i) - F(\underline{n}_{LS}^i - 2|n_S^i, n_L^i, n_R^i)}.$$

Randomization for these two realizations yields a test of exactly size α.

Randomizing over whether to reject a specific null hypothesis might seem unnatural and, indeed, randomized tests are seldom used.[15] In our context, however, we are not interested in whether $p_L^i = p_R^i$ for any particular point game i, but rather whether the null hypothesis is rejected at the expected rate for the thousands of point games

[13] The proof is elementary and omitted.

[14] Suppose $n_{LS}^i \leq \bar{n}_{LS}^i$. The randomized Fisher exact test does not reject the null since $t^i \sim U[F(n_{LS}^i - 1|n_S^i, n_L^i, n_R^i), F(n_{LS}^i|n_S^i, n_L^i, n_R^i)]$ and therefore $t^i \leq F(n_{LS}^i|n_S^i, n_L^i, n_R^i) \leq F(\bar{n}_{LS}^i|n_S^i, n_L^i, n_R^i) \leq \alpha/2$.

[15] See Tocher (1950) for an early general analysis of randomized tests.

in our data set. Our use of the randomized Fisher exact test allows us to exactly control the size of the test, and therefore the expected rejection rate under the null, without any appeal to an asymptotic distribution of a test statistic.

Table 3 shows the percentage of points games for which equality of winning probabilities is rejected for the Hawk-Eye data, for men and women and for both first and second serves. As just noted, for point game i the null hypothesis is rejected at the 5% significance level if either $t^i \leq .025$ or $t^i \geq .975$. Since t^i is random, each percentage is computed for 10,000 trials; the table reports the mean and standard deviation (in parentheses) of these trials. For men, for both first and second serves, the (mean) frequency at which the null is rejected at the 5% significance level is very close to 5%, the level expected if the null is true.[16] For women, the null is rejected at a somewhat higher than expected rate (5.35%) on first serves, and a slightly lower than expected rate (4.86%) for second serves.

		Significance Level	
Setting	# Point Games	5%	10%
Men (1st Serve)	7,198	5.06% (0.16)	10.01% (0.20)
Men (2nd Serve)	7,198	5.02% (0.23)	10.13% (0.30)
Women (1st Serve)	4,108	5.35% (0.22)	10.50% (0.28)
Women (2nd Serve)	4,108	4.86% (0.30)	9.64% (0.40)

Table 3: Rejection Rate (Fisher exact Test) for $H_0 : p_L^i = p_R^i$ (10,000 trials)

At the individual level, the rates at which equality of winning probabilities is rejected at the 5% and 10% significance level are consistent with the theory.

Aggregate Play and the Joint Null Hypothesis

Of primary interest is the joint null hypothesis that $p_L^i = p_R^i$ for each point game i, and we use the t^i's generated from the randomized Fisher exact test to construct our test. Since the t^i's are independent draws from the same continuous distribution, namely the $U[0,1]$ distribution, we can test the joint hypothesis by applying the

[16] Since each point game has fewer second serves than first serves, the stochastic nature of the t's will tend to be more important for second serves. This is evidenced by the higher standard deviations for second serves. Likewise, since we tend to observe fewer serves for women, the standard deviations are higher for women.

Kolmogorov–Smirnov (KS) test to the empirical *c.d.f.* of the t-values. Formally, the KS test is as follows: The hypothesized *c.d.f.* for the t-values is the uniform distribution, $F(x) = x$ for $x \in [0,1]$. The empirical distribution of N t-values, one for each point game, denoted $\hat{F}(x)$, is given by $\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} I_{[0,x]}(t^i)$, where $I_{[0,x]}(t^i) = 1$ if $t^i \leq x$ and $I_{[0,x]}(t^i) = 0$ otherwise. Under the null hypothesis, the test statistic $K = \sqrt{N}\sup_{x\in[0,1]}|\hat{F}(x) - x|$ has a known asymptotic distribution (see p. 509 of Mood, Graybill, and Boes (1974)). Our appeal to an asymptotic distribution at this stage is well justified since there will be thousands of point games for each of the joint null hypotheses we consider.

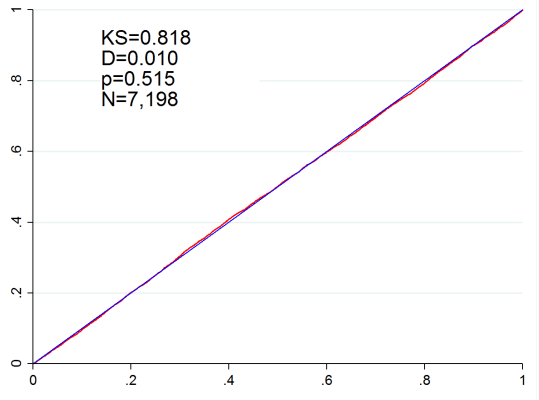
Figure 4(a) shows a realization of an empirical distribution (in red) of t-values for the Hawk-Eye data for first serves by men; the theoretical *c.d.f.* is in blue. Strikingly, the empirical and theoretical *c.d.f.*s very nearly coincide. The value of the KS test statistic is $K = .818$ and the associated p-value is .515. The data is typical of the data that equilibrium play would produce: equilibrium play would generate a value of K at least this large with probability .485. Despite its enormous power, based on 226,298 first serves in 7,198 point games, the test does not come close to rejecting the null hypothesis.

The results reported in Figure 4(a) provide strong support for equilibrium play. Since the t-values are stochastic, the empirical *c.d.f.* and the KS test p-value reported in Figure 4(a) are also random. It is natural to question the robustness of the conclusion that the joint null hypothesis is not rejected to different realizations of the t^i's. To assess its robustness, we run the KS test many times, each time generating a new realization of t-values, a new empirical *c.d.f.* of t-values, a new test statistic K, and a new KS test p-value.

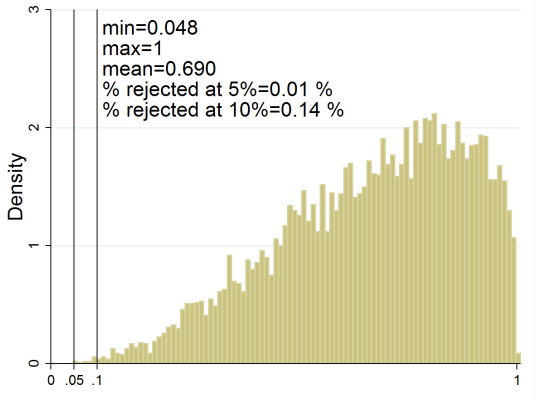
Figure 4(b) shows the empirical density of the KS test p-values obtained after 10,000 repetitions of the test. To construct the density, the horizontal axis is divided into 100 equal-sized bins $[0, .01]$, $[.01, .02], \ldots, [.99, 1.0]$ and so, if 10,000 p-values were equally distributed across bins, then there would be 100 p-values per bin. The vertical height of each bar in the histogram is the number of p-values observed in the bin divided by 100. By construction, the area of the shaded region in Figure 4(b) is one, and hence it is an empirical density. The bins to the left of the vertical lines at .05 and at .10 contain, respectively, p-values for which the null is rejected at the 5% and 10% level.

Figure 4(b) shows that the conclusion above is indeed fully robust to the realiza-

tions of the t-values. The joint null hypothesis of equality of winning probabilities for first serves does not even come close to being rejected for the Hawk-Eye data for first serves by men. In only one instance (.01%) of 10,000 trials of the KS test is the null hypothesis rejected at the 5% level. In only .14% of the trials is it rejected at the 10% level. The mean p-value is .690, which is far from the rejection region.



(a) An empirical *c.d.f.* of t-values



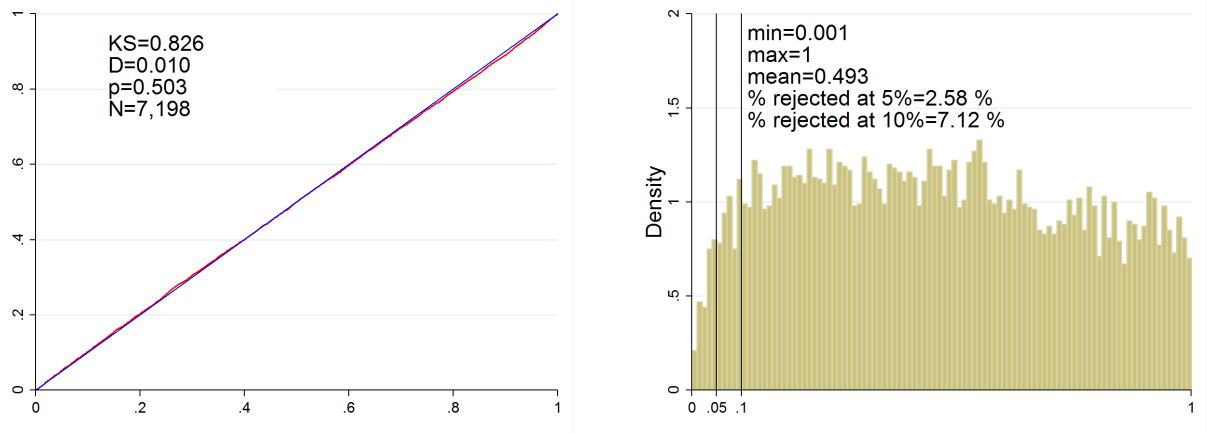
(b) Density of KS test p-values (10,000 trials)

Figure 4: KS test for Men of $H_0 : p_L^i = p_R^i \ \forall i$ (Hawk-Eye, First Serves)

Before proceeding it is important to emphasize several aspects of our test. First, it is a valid test in the sense that if the null hypothesis is true (i.e., $p_L^i = p_R^i$ for each i) then the p-value obtained from the KS test is asymptotically uniformly distributed as the number of point games grows large. Second, conditional on any realization of the data, there is no reason to expect the KS test p-values obtained from running the test repeatedly (e.g., as in Figure 4(b)) to be uniformly distributed. Finally, as the number of serves in each point game grows large, then the intervals $U[F(n_{LS}^i - 1|n_S^i, n_L^i, n_R^i), F(n_{LS}^i|n_S^i, n_L^i, n_R^i)]$ from which the t-values are drawn shrink and the empirical density of the KS p-values collapses to a degenerate distribution.

Figure 5 shows the result of applying our test to the Hawk-Eye data for 86,702 second serves by men from 7,198 point games. For a typical realization of the t-values, such as the one shown in Figure 5(a), the joint null hypothesis of equality of winning probabilities is not rejected. Figure 5(b) shows the density of KS test p-values after 10,000 trials. Only for a small fraction of these trials (2.58%) is the joint null rejected at the 5% level. For second serves as well, the KS test does not come close to rejecting

the joint null hypothesis.



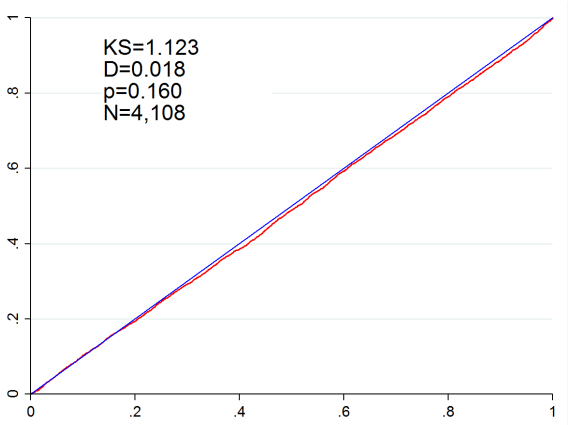
(a) An empirical *c.d.f.* of t-values (b) Density of KS test p-values (10,000 trials)

Figure 5: KS test for Men of $H_0 : p_L^i = p_R^i \ \forall i$ (Hawk-Eye, Second Serves)

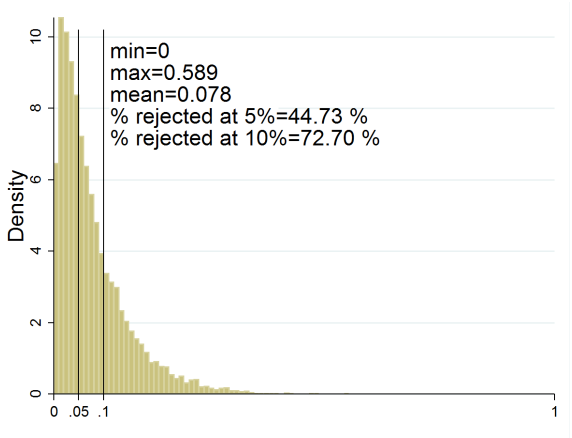
While the data for both first and second serves is strikingly consistent with the theory, comparing Figures 4(b) and 5(b) reveals that for second serves the conformity to theory is slightly less robust to the realization of the t^i's. This is a consequence of the fact that in tennis there are fewer second serves than first serves in each point game. Thus the intervals $U[F(n_{LS}^i - 1|n_S^i, n_L^i, n_R^i), F(n_{LS}^i|n_S^i, n_L^i, n_R^i)]$ from which the t-values are drawn tend to be larger for second serves, and the empirical *c.d.f.* of t-values is more sensitive to the realization of the t's.

Our data also allows a powerful test of whether the play of women conforms to equilibrium. In the Hawk-Eye data for women, there are 110,886 first serves and 41,376 second serves, obtained in 4,108 point games. For women, while the empirical and theoretical *c.d.f.*s of t-values appear to the eye to be close, for many realizations of the t's the distance between them is, in fact, sufficiently large that the joint null hypothesis of equality of winning probabilities is rejected. Figures 6 and 7 show, respectively, the results of KS tests of the hypothesis that $p_L^i = p_R^i$ for all i, for first and second serves. For first serves, the null is rejected at the 5% and 10% significance level in 44.73% and 72.70% of 10,000 trials, respectively. In other words, run once, the test is nearly equally likely to reject the null as not at the 5% level; three out of four times the test rejects the null at the 10% level.

The results for second serves are more ambiguous. The null hypothesis tends not to be rejected at the 5% level: in only 16.01% of the trials is the p-value below .05.

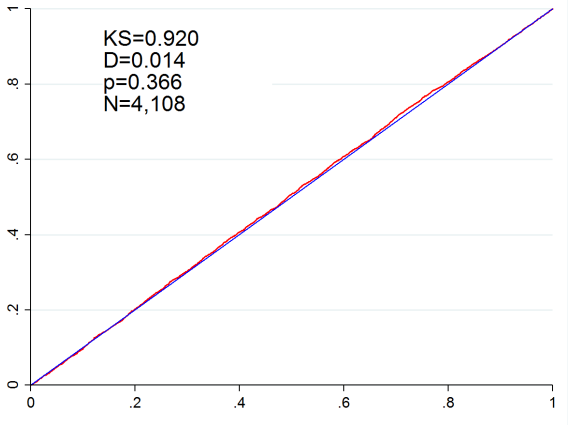


(a) An empirical *c.d.f.* of p-values

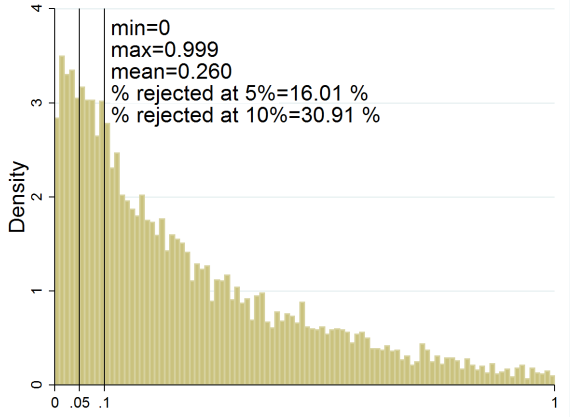


(b) Density KS test p-values (10,000 trials)

Figure 6: KS test for Women of $H_0 : p_L^i = p_R^i \ \forall i$ (Hawk-Eye, First Serves)



(a) An empirical *c.d.f.* of t-values



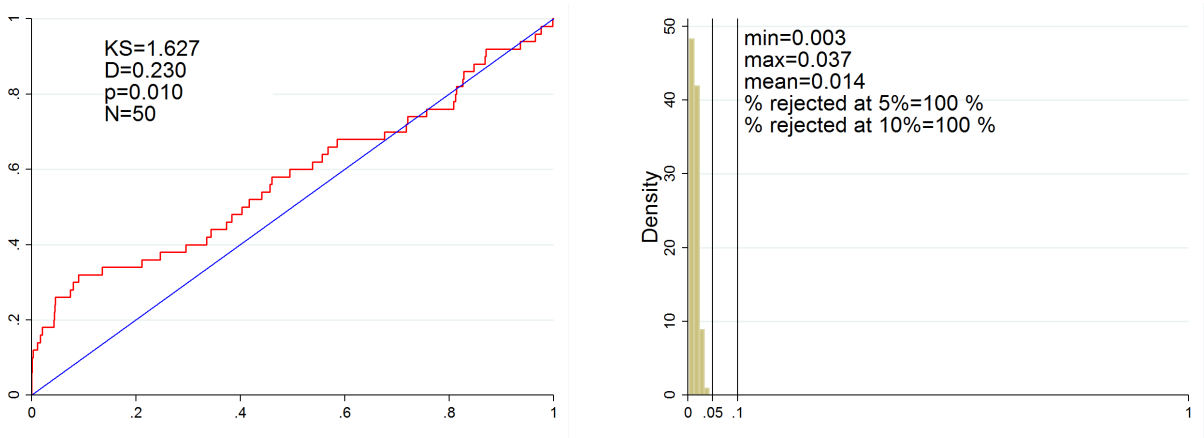
(b) Density of KS test p-values (10,000 trials)

Figure 7: KS test for Women of $H_0 : p_L^i = p_R^i \ \forall i$ (Hawk-Eye, Second Serves)

In sum, male professional tennis players show a striking conformity to the theory on both first and second serves. The behavior of female professional tennis players conforms less closely to the theory, especially on first serves.

The behavior of female professional tennis players, however, conforms far more closely to equilibrium than the behavior of student subjects in comparable laboratory tests of mixed-strategy Nash play. Figure 8(a) shows a representative empirical *c.d.f.* of 50 t-values obtained from applying our test to the data from O'Neill's (1987)

classic experiment in which 50 subjects, in 25 fixed pairs, played a simple card game 105 times. In the game's unique mixed-strategy Nash equilibrium, the probability that player i wins a hand is the same when playing the joker card as when playing a number card (i.e., $p_J^i = p_N^i$). Nonetheless, for the empirical *c.d.f.* of t-values in Figure 8(a), the joint null hypothesis of equality of winning probabilities is decisively rejected, with a p-value of .01. The empirical density function in Figure 8(b) shows that rejection of the null at the 5% significance level is completely robust to the realization of the t-values – at this significance level the null is certain to be rejected.



(a) An empirical *c.d.f.* of t-values (b) Density of KS test p-values (10,000 trials)

Figure 8: KS test of $H_0 : p_J^i = p_N^i \ \forall i$ (O'Neill's (1987) experimental data)

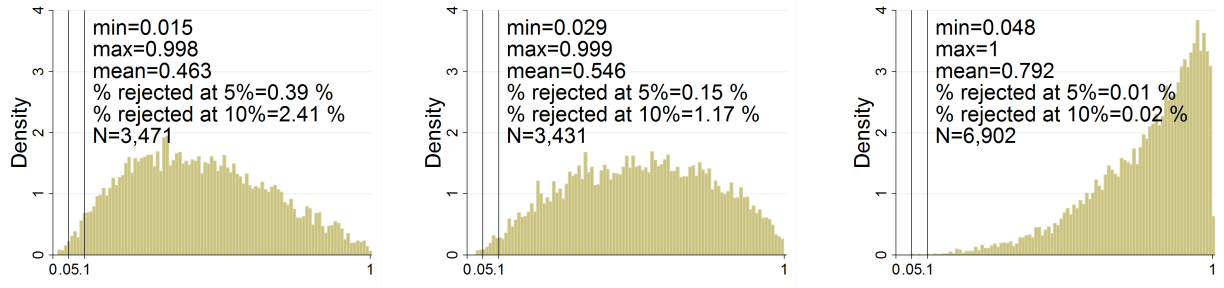
Hence, the behavior of female professional tennis players conforms far more closely to theory than the behavior of student subjects in laboratory experiments.

Expertise and Equality of Winning Probabilities

Next we consider whether the behavior of better (i.e., higher ranked) players conforms more closely to theory. The ATP (Association of Tennis Professionals) and the WTA (Women's Tennis Association) provide rankings for male and female players, respectively. Our analysis here is based on the subsample of matches for which we were able to obtain the receiver's rank at the time of the match. It consists of 96% of all point games for men but, since the ranking data was unavailable for women for the years 2005 and 2006, only 69% of the point games for women.[17] The median rank for male players is 22 and for female players is 17.

[17]The ATP/WTA ranking were obtained from http://www.tennis-data.co.uk/alldata.php.

In a mixed-strategy equilibrium the *receiver*'s play equalizes the server's winning probabilities. (In the illustrative example in Figure 2, the servers's winning probability is .65 for each of his actions when the receiver follows his equilibrium mixture.) Thus to evaluate the effect of expertise on behavior, we partition the data for first serves into two subsamples based on whether the rank of the player receiving the serve was above or below the median rank. We say players with a median or higher rank are "top" players; all other players are "non-top." The three panels of Figure 9 show the empirical *c.d.f.*s of KS-test p-values when testing the joint null hypothesis of equality of winning probabilities for each subsample of men and for the sample of all point games for which we could obtain the receiver's rank. Panel (c) is the analogue of Figure 4(b) and shows that the null of hypothesis that winning probabilities are equalized is not rejected for the sample of 6,902 point games for which we have the receiver's rank.



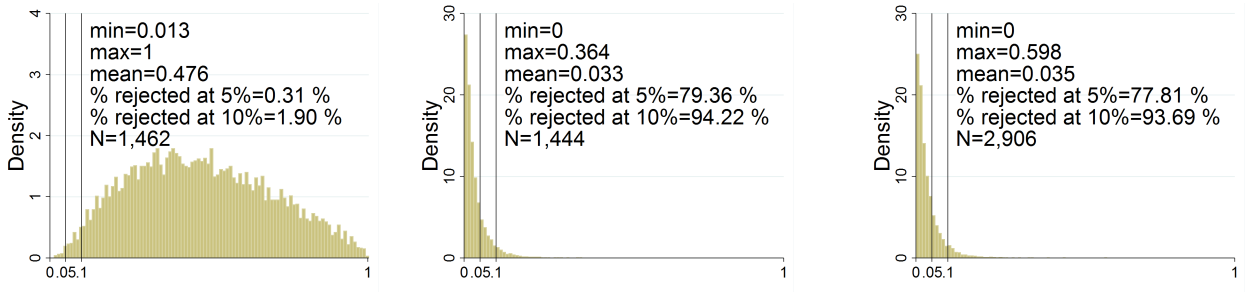
(a) "Top" male receivers (b) "Non-top" male receivers (c) All male receivers

Figure 9: KS test for Men of $H_0 : p_L^i = p_R^i \ \forall i$ by receiver's rank

Panels (a) and (b) of Figure 9 show that the null hypothesis that winning probabilities are equalized is not rejected when servers face either top or non-top male receivers. In only .39% and .15% of the trials is the null rejected at the 5% significance level; p-values are typically far from the rejection region. Hence we do not come close to rejecting the hypothesis that male receivers act to equalize the server's winning probability, for either top or non-top receivers. This result is not surprising given the close conformity of the data to the theory on the whole sample.

Figures 6 and 7 established that the behavior of female professional tennis players conformed less neatly to equilibrium. For first serves, the joint null hypothesis of equality of winning probabilities is rejected at the 5% level in 44.73% of all trials. Figure 10(c) shows that a similar conclusion holds for the subsample of 2,906 point

games for which we were able to obtain the receiver's rank.



(a) "Top" female receivers (b) "Non-top" female receivers (c) All female receivers

Figure 10: KS test for Women of $H_0 : p_L^i = p_R^i \ \forall i$ by receiver's rank

Figures 10(a) and (b) show a striking difference between the play of top and non-top female receivers: for the subsample of matches in which the receiver is ranked "top" the joint null hypothesis of equality of winning probabilities does not come close to being rejected, as shown in panel (a). In contrast, the null is decisively rejected for the subsample in which the receiver is ranked "non-top," as shown in panel (b). The best female players, when receiving the serve, do act to equalize the server's winning probabilities, in accordance with equilibrium. To our knowledge, this is the first evidence in the literature showing that the behavior of better players conforms more closely to theory.

Why do both top and non-top male receivers equalize the server's winning probabilities, while for women only top receivers do? We conjecture that the selection pressure towards equilibrium is larger for men than for women. A male receiver who fails to equalize the server's winning probabilities can readily be exploited since men deliver serves at very high speed. Such a receiver may well not be sufficiently successful to appear in our dataset. The serve in women's tennis, by contrast, is much slower – fewer serves are won by aces and the serve is more frequently broken. Return and volley play is relatively more important in women's tennis, and a good return and volley player can be successful even if her play when receiving a serve is somewhat exploitable. We find that the best female players, nonetheless, do equalize the server's winning probabilities.

Comparing tests

We conclude this section by discussing the differences between our test and the WW test of the joint null hypothesis that winning probabilities are equalized. Our test is based on the empirical *c.d.f.* of the t-values obtained from the randomized Fisher exact test, which are exactly distributed $U[0,1]$ under the null hypothesis that $p_L^i = p_R^i$ for each point game i. WW's test is based on the empirical *c.d.f.* of the Pearson goodness of fit p-values, which are only asymptotically distributed $U[0,1]$ under the null. In particular, for each point game i, WW compute the test statistic

$$Q^i = \sum_{j \in \{L,R\}} \left[\frac{(n_{jS}^i - n_j^i \hat{p}^i)^2}{n_j^i \hat{p}^i} + \frac{(n_{jF}^i - n_j^i (1-\hat{p}^i))^2}{n_j^i (1-\hat{p}^i)} \right],$$

where $\hat{p}^i = (n_{LS}^i + n_{RS}^i)/(n_L^i + n_R^i)$, the server's empirical win rate, is the maximum likelihood estimate of the true but unknown winning probability p^i. The test statistic Q^i is asymptotically distributed chi-square with 1 degree of freedom under the null hypothesis as the number of serves grows large, and the associated p-value is therefore only asymptotically distributed $U[0,1]$. The p-value is not exactly distributed $U[0,1]$ for any finite number of serves and thus, when the number of point games grows large relative to the number of serves in each point game, the WW test rejects the joint null hypothesis even when it is true.

In Section 6 we verify via Monte Carlo simulations that the WW test is not valid when the number of point games is large relative to the number of serves. We show it is valid when the number of points games is not large relative to the number of serves, at it was for the WW data set of 40 points games.[18] Monte Carlo simulations show that our test is more powerful than the WW test when the WW test is valid. Our test, therefore, has two significant advantages over the WW test: (i) it is valid even when the number of point games is large, and (ii) it is more powerful than the WW test.

[18] Whether the WW test is valid for any particular sample can be determined by Monte Carlo simulations. In favor of the WW test, it makes a deterministic decision – it either rejects the null or not – and hence the results are easier to interpret, even if it falsely rejects a true null when the number of point games is large.

5 Serial Independence

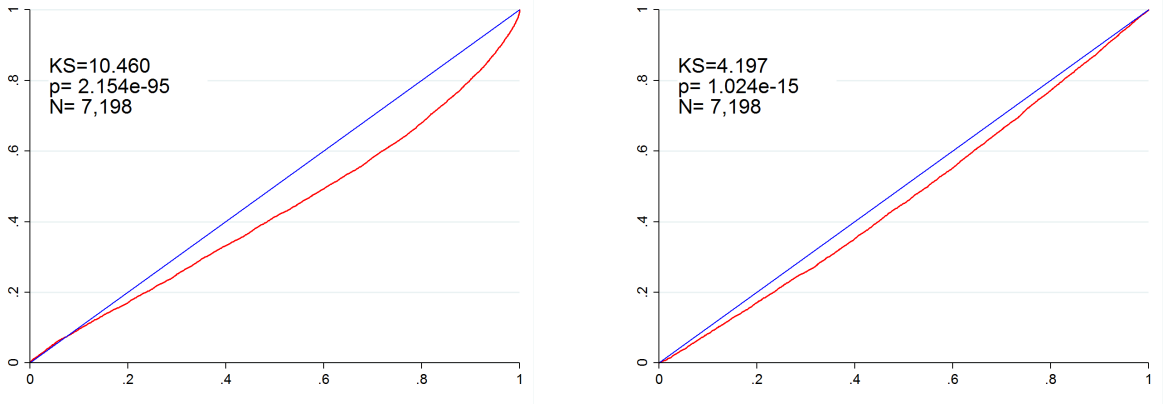
We test the hypothesis that the server's choice of direction of serve is serial independent. For each point game i, let $s^i = (s^i_1, \ldots, s^i_{n^i_L+n^i_R})$ be the sequence of first-serve directions, in the order in which they occurred, where $s^i_n \in \{L, R\}$. Let r^i denote the number of runs in s^i. (A run is a maximal string of identical symbols, either all L's or all R's.) Under the null hypothesis of serial independence, the probability that there are exactly r runs in a randomly ordered list of n^i_L occurrences of L and n^i_R occurrences of R is known. Denote this probability by $f_R(r|n^i_L, n^i_R)$ and let $F_R(r|n^i_L, n^i_R)$ be the associated *c.d.f.* At the 5% significance level, the null is rejected if $F_R(r|n^i_L, n^i_R) \leq .025$ or if $1 - F_R(r-1|n^i_L, n^i_R) \leq .025$, i.e., if the probability of r or fewer runs is less than .025 or the probability of r or more runs less less than .025. In the former case, the null is rejected since there are too few runs, i.e., the server switches the direction of serve too infrequently to be consistent with randomness. In the later case, the null is rejected as the server switches direction too frequently.

To test the joint null hypothesis that first serves are serially independent, for each point game i we draw the random test statistic t^i from the $U[F_R(r^i-1|n^i_L, n^i_R), F_R(r^i|n^i_L, n^i_R)]$ distribution. Under the joint null hypothesis of serial independence, each t^i is distributed $U[0, 1]$. We then apply the KS test to the empirical distribution of the t-values.

Figure 11 shows representative empirical *c.d.f.*s of t-values for first serves (left panel) and for second serves (right panel) for the Hawk-Eye data for men. The KS test rejects the joint null hypothesis of serial independence, for both first and second serves, with p-values virtually equal to zero.[19] In each case, the empirical *c.d.f.* lies below the theoretical *c.d.f.*, and hence the null is rejected as a consequence of too

[19] At the individual player level, serial independence is rejected in point game i at the 5% significance level if $t^i \leq .025$ or $t^i \geq .975$. For first serves, we reject serial independence as a result too few runs (i.e., $t^i \leq .025$) for 2.9% of the point games, and reject it as a result of too many runs (i.e., $t^i \geq .975$) for 7.0% of the point games.

much switching, i.e., there are more than the expected number of large t-values.

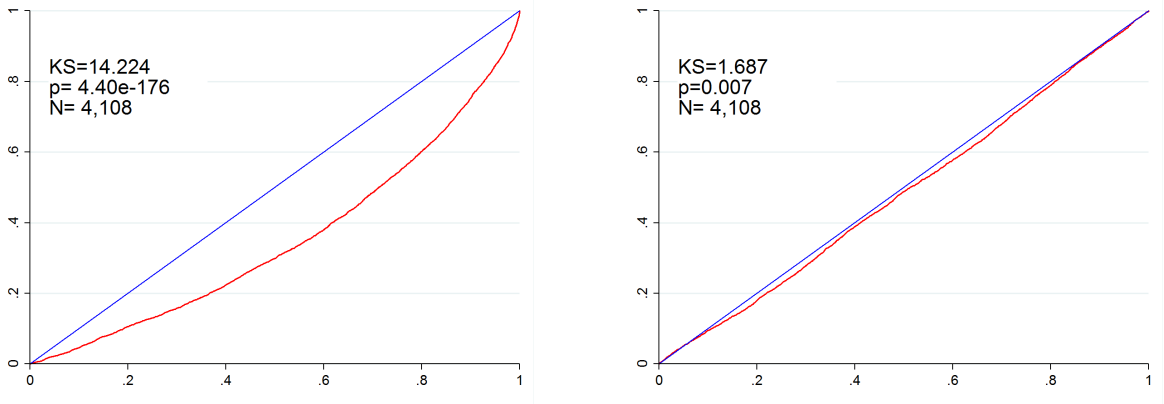


(a) First Serve: Empirical *c.d.f.* of t-values (b) Second Serve: Empirical *c.d.f.* of t-values

Figure 11: KS test for Men of $H_0 : s^i$ is serial independent $\forall i$ (Hawk-Eye)

The empirical density of the KS test p-values that we have provided in prior figures is omitted for Figure 11 since the p-values are virtually zero for every realization of the t's.

Figure 12 shows representative empirical *c.d.f.*s of t-values for first and second serves by women for the Hawk-Eye data. Women also exhibit negative serial correlation in the direction of serve, for both first and second serves, with the null of serial independence rejected at virtually any significance level.



(a) First Serve: Empirical *c.d.f.* of t-values (b) Second Serve: Empirical *c.d.f.* of t-values

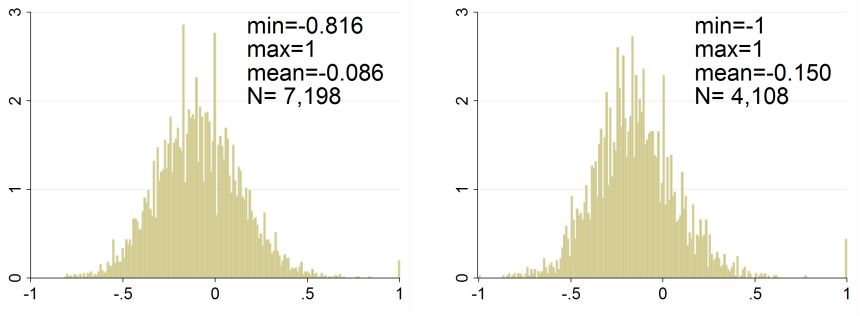
Figure 12: KS test for Women of $H_0 : s^i$ is serial independent $\forall i$ (Hawk-Eye)

Comparing Figures 11(a) and 12(a) one might be tempted to conclude that women exhibit more serial correlation in first serves than men since the empirical *c.d.f.* of

t-values is further from the theoretical one (*viz.* the 45-degree line) for women. While this conclusion is correct, as we shall see shortly, it is premature: when the server's choice of direction of serve is not serially independent in point game i, then the distribution of t^i will tend to depend on the number of first serves. Since we observe different numbers of first serves for men and women and, indeed, different numbers of first serves for different players, a direct comparison of the *c.d.f.*s is not meaningful.

Gender and Serial Correlation

To determine the degree of serial correlation in first serves, and whether the difference between male and female players is statistically significant, we compute, for every point game, the Pearson product-moment correlation coefficient between successive serves.[20] Figure 13 shows the empirical densities of correlation coefficients for male and female tennis players for first serves.



(a) Men (Hawk-Eye) (b) Women (Hawk-Eye)

Figure 13: Empirical density of correlation coefficients, first serves

The mean correlation coefficient for men is -0.086 and for women is -0.150, a statistically significant difference using a two-sample t-test.[21]

Table 4 shows the result of a logit regression for first serves in which the dependent variable is the direction of the current serve and the independent variables are the direction of the prior serve (from the same point game) and the direction of the prior serve interacted with gender. We use a fixed effect logit, using only within point game

[20] When all serves are in the same direction we take the correlation coefficient to be one.

[21] The two-sample t-test yields a test statistic of -14.16 and p-value of 4.57×10^{-39}.

variation, to cancel out variation in the equilibrium mixture across point games.[22]

$Right_{t-1}$	-0.659
	$(p < 0.001)$
$Right_{t-1} \times male$	0.329
	$(p < 0.001)$
N_{serves}	$325,394$
Fixed Effect	point game

Table 4: Serial Correlation and Gender

The coefficient estimate on $Right_{t-1} \times male$ is statistically significant and positive, indicating that men exhibit less negative serial correlation in their choices than women. The estimated magnitude of serial correlation is strategically significant. To illustrate, consider a female player who (unconditionally) serves right and left with equal probability. If the prior serve was right, the estimates predict that the next serve will be right with probability 0.418 if the server is male but will be right with probability only 0.341 if the server is female.

Expertise and Serial Correlation

We now provide additional evidence that the behavior of better players conforms more closely to optimal and equilibrium play. Optimal play for the server requires that the direction of serve be serially independent, since serial correlated play is predictable and therefore exploitable. Here we show that higher ranked male players exhibit less serial correlation than lower ranked players, while the degree of serial correlation does not depend on rank for women. This provides additional evidence, consistent with our earlier conjecture, that there is a strong selection effect against men who depart from equilibrium play.[23]

Table 5 shows the results of logit regressions in which the dependent variable is the direction of the current first serve and the independent variables are the direction of the prior first serve (in the same point game), the direction of the prior serve interacted with the server's rank, and the direction of the prior serve interacted

[22] Estimating the fixed effect logit regression requires that point games in which all first serves are in the same direction be dropped from the sample.

[23] In Section 4 we focused on the ranks of receivers since it is the receiver's mixture that determines whether server's winning probabilities are equalized.

with the receiver's rank. We measure rank as proposed by Klaassen and Magnus (2001), transforming the ATP/WTA rank of a player into the variable $\tilde{R}$ where $\tilde{R} = 8 - \log_2(\text{ATP/WTA rank})$. Higher ranked players have *higher* values of $\tilde{R}$, e.g., the players ranked first, second, and third have values of $\tilde{R}$ equal to 8, 7, and 6.415, respectively.

	Men	Women
$Right_{t-1}$	$-.577$	$-.689$
	$(p < 0.001)$	$(p < 0.001)$
$Right_{t-1} \times \tilde{R}_{server}$	0.067	0.008
	$(p < 0.001)$	$(p = 0.359)$
$Right_{t-1} \times \tilde{R}_{receiver}$	-0.002	0.004
	$(p = 0.628)$	$(p = 0.610)$
N_{serve}	$207,418$	$77,508$
$N_{pointgame}$	$6,887$	$2,901$
Fixed Effect	point game	point game

Table 5: Serial Correlation and Player Rank

For men, the coefficient on $Right_{t-1} \times \tilde{R}_{server}$ is positive and statistically significant. Men exhibit less correlation in their direction of serve as they are more highly ranked. For women, by contrast, the server's rank is statistically insignificant. As expected, the rank of the receiver is statistically insignificant for both men and women.

6 Statistical Power and a Re-evaluation

In this section we use Monte Carlo simulations to study the properties of the KS test of the joint hypothesis of equality of winning probabilities. We show that the test is valid when the empirical *c.d.f.* is generated from the Pearson goodness of fit test p-values, so long as the number of point games is not too large (as it was in WW). If, however, the number of points games is large, then the same test rejects the null even when it is true, and is thus not valid. We show, in contrast, that if the empirical *c.d.f.* is generated from the randomized Fisher exact test t-values as we propose, then the test is valid even when the number of point games is large.

We show further that our KS test based on the randomized Fisher t-values is more powerful than the two tests used in the prior literature: the KS test based on the Pearson goodness of fit p-values and the Pearson joint test.[24] A more powerful test has the potential to reverse the conclusions from WW (for men) and HHT (for men, women, and juniors) that the win rates in the serve and return play of professional tennis players are consistent with equilibrium. We show that the conclusion of WW and HHT for men are robust. However, the more powerful test does not support HHT's finding that the serve and return play of female professional tennis players and of players in junior matches is consistent with theory.

The Power of Our Test

To evaluate the power of the KS test based on the randomized Fisher exact test t-values, we follow WW and frame our discussion in terms of the hypothetical point game in Figure 2. Recall that in the game's mixed-strategy Nash equilibrium, the receiver chooses L with probability 2/3. Denote by θ the probability that the receiver chooses L. Our null hypothesis H_0 that $p_L = p_R$ can equivalently be viewed as the null hypothesis that $\theta = 2/3$, i.e., the receiver follows his equilibrium mixture, thereby equalizing the server's winning probabilities. Denote by $H_a(\theta)$ the alternative hypothesis that the receiver chooses L with probability θ. Then the server's winning probabilities are

$$p_L(\theta) = .58\theta + .79(1 - \theta)$$

and

$$p_R(\theta) = .73\theta + .49(1 - \theta).$$

We conduct Monte Carlo simulations to compare the power of our test, i.e., the probability that H_0 is rejected when $H_a(\theta)$ is true, to the tests used in the prior literature.

Since we have data for 7198 points games for men, we first simulate data for 7000 points games with payoffs as given above. In the simulated data every point game has 70 serves, and serves in each direction are equally likely.[25] Table 6 shows, as θ varies near its equilibrium value of 2/3, the probability that the joint null hypothesis

[24] See, for example, Table 1 and Figure 2 in WW.

[25] Simulating the data with the hypothetical point game's 8/15 equilibrium mixture probability on left has no impact on the results.

$H_0 : p_L^i = p_R^i \ \forall i \in \{1, \ldots, 7000\}$ is rejected at the 5% significance level when $H_a(\theta)$: $p_L^i = p_L(\theta)$ and $p_R^i = p_R(\theta) \ \forall i \in \{1, \ldots, 7000\}$ is true, for several different tests.

True θ	KS based on t's	KS based on p's	Pearson joint test
0.65	0.997	0.581	0.293
0.66	0.460	0.541	0.226
2/3	0.046	0.527	0.213
0.67	0.153	0.529	0.213
0.68	0.964	0.558	0.268

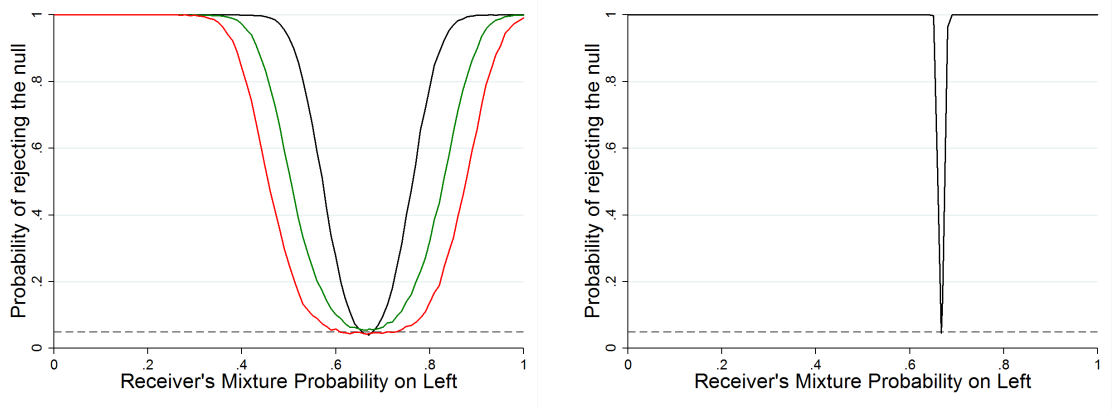
Table 6: Rejection rate for H_0 at the 5% level, $N = 7000$

The first column of Table 6 shows the probability of rejecting the null when using the KS test based on the randomized Fisher exact test t-values. Note that the test is valid. Specifically, if the null is true, i.e., $\theta = 2/3$, then the null is rejected with probability approximately 0.05. By contrast, if the null is false, e.g., $H_a(.65)$ is true, i.e., the server's true winning probability is $p_L(.65) = .6535$ for serves left and $p_R(.65) = 0.6460$ for serves right, then H_0 is rejected at the 5% level with probability .997. The second column of Table 6 shows that the KS test based on the Pearson goodness of fit p-values is not valid: it rejects the null hypothesis at the 5% significance level with probability 0.527 when the null is true. The third column shows that the Pearson joint test (see WW p. 1527 for a description of this test) is also not valid.

Figure 14(b) shows the power of KS test based on the randomized Fisher exact test t-values for all values of θ. It shows that our test, coupled with a large dataset, yields an extraordinarily powerful test of the joint null hypothesis of equality of winning probabilities. The power functions for the Pearson joint test and the KS test based on the p-values from the Pearson goodness of fit test are omitted since, as shown in Table 6, neither test is valid.

Figure 14(a) compares the power of the three tests discussed above for a sample size of 40 point games, the number of point games in WW's dataset. It shows that the probability that the joint null hypothesis $H_0 : p_L^i = p_R^i \ \forall i \in \{1, \ldots, 40\}$ is rejected at the 5% significance level when $H_a(\theta) : p_L^i = p_L(\theta)$ and $p_R^i = p_R(\theta) \ \forall i \in \{1, \ldots, 40\}$ is true. The power function in red (the curve at bottom) shows the probability of rejecting H_0 when $H_a(\theta)$ is true for the KS test based on the empirical distribution

of the 40 p-values from the Pearson goodness of fit test.[26] Importantly, it shows that this test is valid for the WW sample. The power function in green (middle curve) is for the Pearson joint test, and is the analogue of the power function shown in Figure 4 of WW. The power function in black (the curve at top) is for the KS test based on the empirical distribution of 40 t-values from the randomized Fisher exact test. This last test is, by far, the most powerful. If, for example, $H_a(.6)$ is true, then the KS test based on the t's rejects H_0 at the 5% significance level with probability 0.273, while the Pearson joint test and the KS test based on the Pearson goodness of fit p-values reject H_0 with probability 0.101 and 0.057, respectively.



(a) $N = 40$ (b) $N = 7000$

Figure 14: Power Functions for KS test based on t-values (black), p-values (red), and Pearson joint (green)

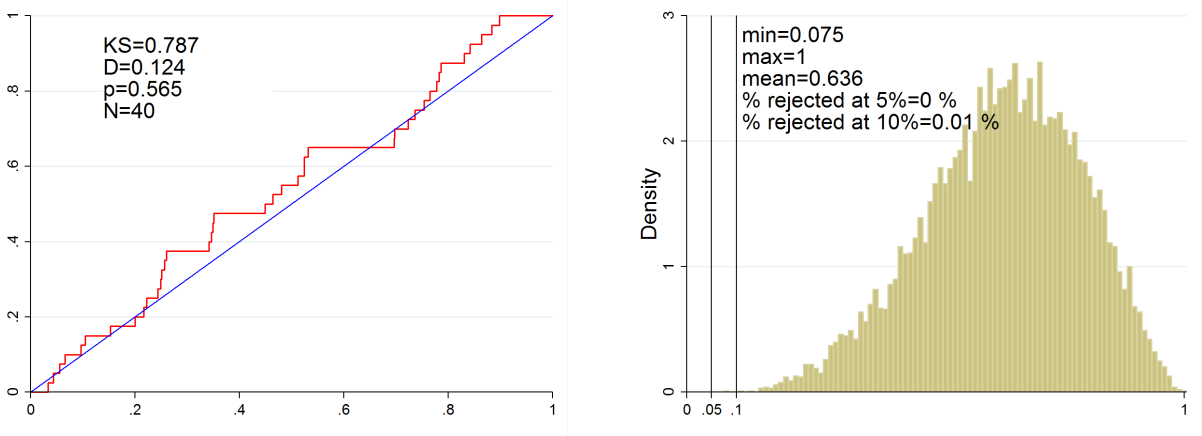
Re-Analysis of Prior Findings

The KS test we propose, based on the randomized Fisher exact test t-values, is valid for all sample sizes and is more powerful than the existing tests used in the literature. Given its greater power, our test has the potential to overturn results in the prior literature based on less-powerful tests.

Using the KS test based on the Pearson goodness of fit p-values, WW found that the joint null hypothesis of equality of winning probabilities did not come close to being rejected. Figure 15(a) shows one realization of the empirical distribution of Fisher exact test t-values for the WW data. For this realization, the value of the test

[26] For the power functions reported in Table 6 and Figure 14, the data is simulated 10,000 times for each value of $\theta \in \{0, .01, .02, \ldots, .99, 1\}$ and for $\theta = 2/3$.

statistic is $K = .787$ and the associated p-value is .565. Figure 15(b) shows that the KS test p-values after 10,000 trials are concentrated around .6, and hence are far from the rejection region. The joint null hypothesis of equality of winning probabilities is not rejected once at the 5% significance level. Thus the new test confirms the WW finding that the joint null hypothesis of equality of winning probabilities for first serves does not come close to being rejected for male professional tennis players.



(a) An empirical *c.d.f.* of t-values (b) Density of KS test p-values (10,000 trials)

Figure 15: KS test of $H_0 : p_L^i = p_R^i \ \forall i$ (WW data)

HHT studies a dataset comprised of ten men's matches, nine women's matches, and eight junior's matches. The men's and women's matches are all from Grand Slam finals, while the juniors matches include the finals, quarterfinals, and second-round matches in both tournaments and Grand Slam matches. HHT found, using the KS test based on Pearson p-values, that the joint null hypothesis of equality of winning probabilities is not rejected for any one of their datasets, or all three jointly. The KS statistics are 0.778 for men (p-value .580), 0.577 for women (p-value .893), 0.646 for juniors (p-value .798), and 0.753 (p-value .622) for all 27 matches or 108 point games combined. We show that this conclusion is robust for men, but not for women and juniors, to using the more powerful test based on the t-values.

Figure 16(a) shows, for the HHT men's data, a representative empirical *c.d.f.* of t-values (left panel) and the empirical distribution of the p-values (right panel) obtained from 10,000 trials of the KS test based on the randomized Fisher t-values. The joint null hypothesis is not rejected once at the 5% level. Hence the more powerful test

supports HHT's findings for men.

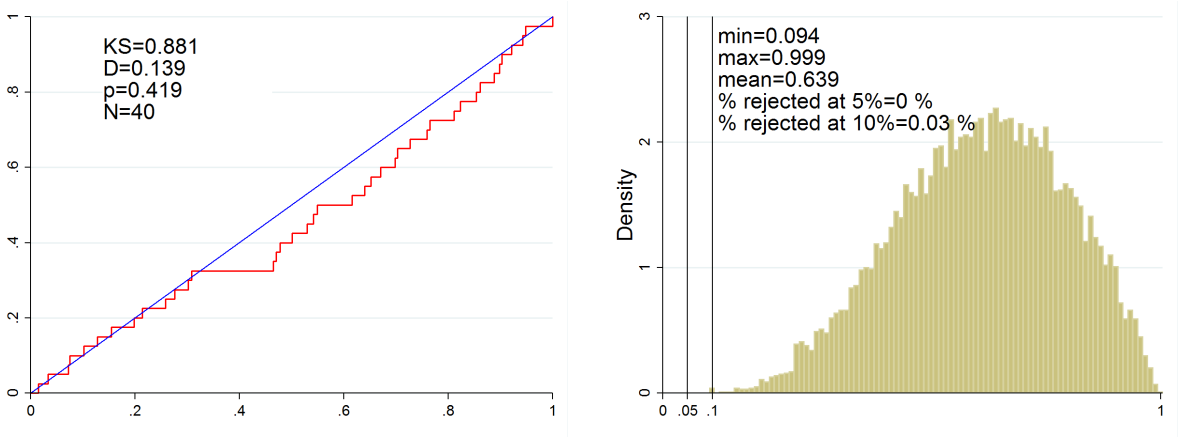


Figure 16(a): KS test for Men of $H_0 : p_L^i = p_R^i \ \forall i$ (HHT data)

Figures 16(b) and (c) show for women and juniors, by contrast, the empirical distributions of p-values are shifted sharply leftward (relative to the one for men) and the same joint null hypothesis is frequently rejected. For women, for example, it is rejected in 18.13% of 10,000 trials at the 5% level and in 47% of all trials at the 10% level. The leftward shift of the empirical density of the p-values is even more striking for juniors. For that data, the joint null is rejected at the 5% level in 49.18% of the trials and at the 10% level in 77.28% of the trials.

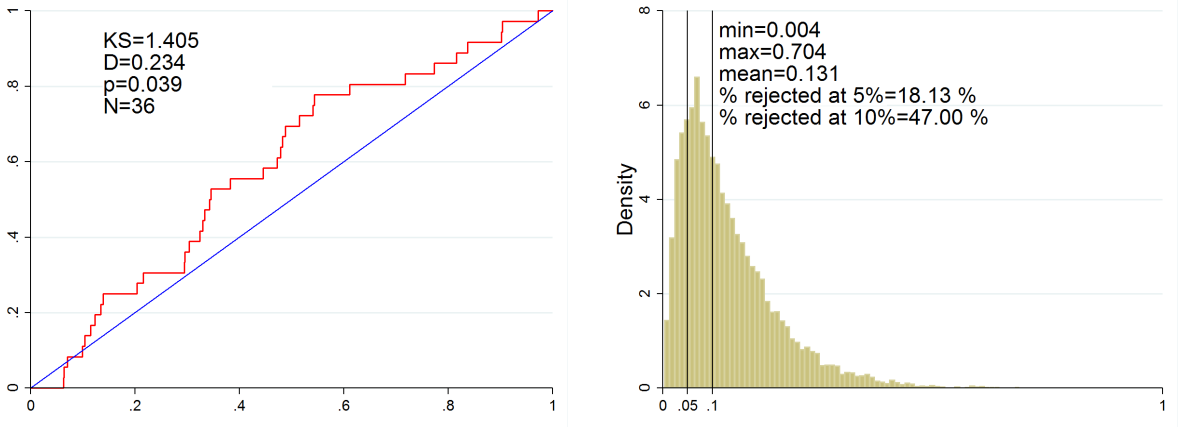


Figure 16(b): KS test for Women of $H_0 : p_L^i = p_R^i \ \forall i$ (HHT data)

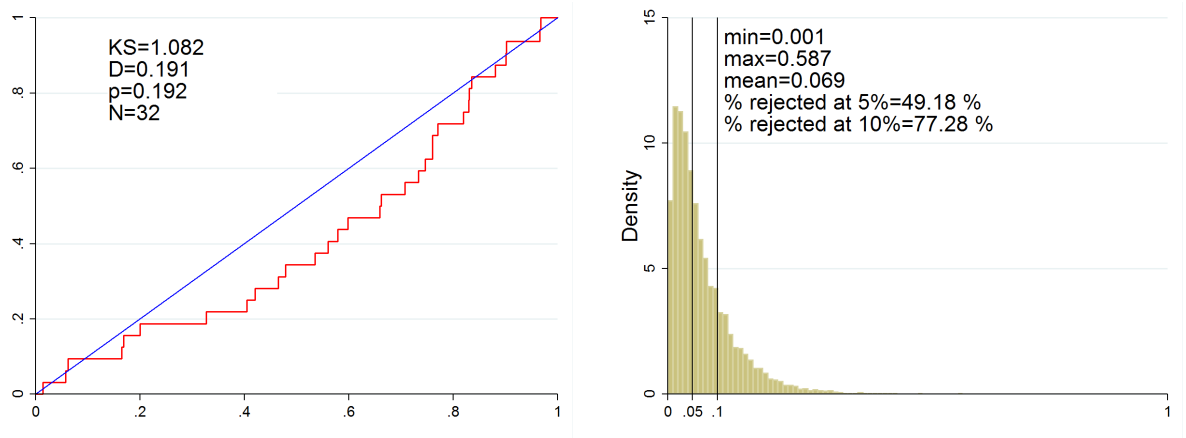


Figure 16(c): KS test for Juniors of $H_0 : p_L^i = p_R^i \ \forall i$ (HHT data)

Thus the greater power of the KS test based on the t-values confirms HHT's conclusion for men, but overturns their conclusions for women and juniors.

7 Conclusion

Data from professional tennis is uniquely well suited to test the empirical validity of game theory: Players are highly experienced and are well incentivized, and there is high quality data on the players' abilities, their actions, and the outcomes. Theory generates clear testable predictions. The results reported here provide strong evidence that winning probabilities are equalized across directions of serve, just as theory predicts. These results provide support for the empirical relevance outside of the laboratory, not only for mixed-strategy Nash equilibrium, but for game theory generally.

Our results on serial correlation show that the behavior of even highly experienced players may fail to conform with theory when it is difficult to identify or exploit departures from equilibrium play, and so selection pressures are low. We find evidence that even among experienced players, the behavior of better players conforms more closely to equilibrium.

8 Appendix A: Data Cleaning

There were several steps in the cleaning the data. The table below shows the numbers of serves remain after each step. As noted in the text, we first eliminated from our analysis every game in which the scoreline did not evolve logically. Row (i) shows the number of first serves, second serves, and point games that remain. We then eliminated those serves in which there is ambiguity regarding which player is serving (Row (ii)), and serves in which there is ambiguity regarding whether the serve is a first or second serve (Row (iii)). Finally, if a point game has fewer than 10 first serves, then we drop the point game and also the associated point game of second serves (Row (iv)).[27]

		Female			Male		
		1^{st}	2^{nd}	N	1^{st}	2^{nd}	N
	All	147,000	57,005	4,657	284,109	113,757	7,951
(i)	Scoreline	115,014	44,082	4,511	230,305	91,341	7,690
(ii)	Server?	113,125	43,387	4,511	228,802	90,739	7,690
(iii)	1^{st} or 2^{nd}?	113,121	42,180	4,511	228,785	87,732	7,690
(iv)	≥ 10 serves	110,886	41,376	4,108	226,298	86,702	7,198

Table A1: Number of serves and point games after data cleaning.

9 Appendix B: A Deterministic Test

Here we construct a deterministic test of the joint null hypothesis that winning probabilities are equal for serves to the left and serves to the right. For the test we construct, the probability under the null of each realization of the data is known, and hence the distribution of the test statistic under the null can be determined via Monte Carlos simulations. While the test is valid (by construction), we show that it has substantially less statistical power than the KS test based on the randomized Fisher exact test t values that we develop and employ in the body of the paper. The lower statistical power of the determinist test means that it sometimes fails to reject a null hypothesis when the same null is rejected with the randomized Fisher exact test, as we show below.

[27] Our results are robust to the choice of restrictions (e.g., more than 10, 20, or 30 serves).

THE TEST

We test the joint null hypothesis that $p_L^i = p_R^i$ for each point game $i \in \{1, \ldots, N\}$.[28] The test is based on the empirical winning probabilities for serves left and serves right, denoted for point game i by $\hat{p}_L^i = n_{SL}^i / n_L^i$ and $\hat{p}_R^i = n_{SR}^i / n_R^i$, respectively. In particular, let $\hat{X} = \sum_{i=1}^N |\hat{p}_L^i - \hat{p}_R^i|$ be the sum of the absolute differences of the empirical winning probabilities over N point games. Let $n_S = (n_S^1, \ldots, n_S^N)$, $n_L = (n_L^1, \ldots, n_L^N)$, and $n_R = (n_R^1, \ldots, n_R^N)$ be the marginals for the N point games. Let $F_{|n_S, n_L, n_R}(X)$ denote the distribution of X, conditional on the marginals, under the null hypothesis. Then the two-sided p-value for $\hat{X}$ is

$$p = 2 \min(F_{|n_S, n_L, n_R}(\hat{X}), 1 - F_{|n_S, n_L, n_R}(\hat{X})).$$

Since $F_{|n_S, n_L, n_R}(X)$ is unknown, we obtain p-values by generating B simulated values of the test statistic under the null hypothesis. Recall that in point game i the probability of k winning serves to the left, under the null hypothesis that $p_L^i = p_R^i$ and given the marginals n_S^i, n_L^i, and n_R^i, is given by

$$f(k | n_S^i, n_L^i, n_R^i) = \frac{\binom{n_L^i}{k} \binom{n_R^i}{n_{RS}^i}}{\binom{n_L^i + n_R^i}{n_S^i}},$$

where $n_{RS}^i = n_S^i - k$. Hence $f(k | n_S^i, n_L^i, n_R^i)$ is the probability that $\hat{p}_L^i = k / n_L^i$ and $\hat{p}_R^i = (n_S^i - k) / n_R^i$ are the empirical winning probabilities for serves left and serves right. Simulating empirical winning probabilities for each of the N point games according to these distributions, we obtain a simulated value for the test statistic. Let $\hat{X}^*_{|n_S, n_L, n_R}(j)$ denote the j-th simulated value of the test statistic. Following MacKinnon (2009), the equal-tailed simulated p-value for $\hat{X}$ is then[29]

$$p^* = 2 \min \left(\frac{1}{B} \sum_{j=1}^B I(\hat{X}^*_{|n_S, n_L, n_R}(j) \leq \hat{X}), \frac{1}{B} \sum_{j=1}^B I(\hat{X}^*_{|n_S, n_L, n_R}(j) > \hat{X}) \right).$$

Table B1 shows the test statistics, with the associated simulated p-values and sample sizes, for male and female players and first and second serves, of the test of

[28] We drop point games in which either $n_L^i = 0$ or $n_R^i = 0$. By comparison, our KS test based on the randomized Fisher t^i values calls for a draw $t^i \sim U[0, 1]$ for such point games, which reflects that they are not informative about the null.

[29] MacKinnon (2009) calls this the equal-tailed "bootstrap" p-value.

the null hypothesis that winning probabilities are equalized.

		Men		Women	
Sample		First Serve	Second Serve	First Serve	Second Serve
All	$\hat{X}$	1090.5	1890.7	704.5	1098.9*
	p^*	0.613	0.540	0.252	0.024
	N	7,188	6,131	4,095	3,483
With Ranking	$\hat{X}$	1045.9	1813.5	489.0	784.0
	p^*	0.902	0.618	0.493	0.145
	N	6,892	5,856	2,902	2,486
Top Receiver	$\hat{X}$	539.2	915.4	244.9	385.9
	p^*	0.240	0.536	0.502	0.362
	N	3,462	2,926	1,461	1,254
Non-top Receiver	$\hat{X}$	506.6	898.0	244.1	398.1
	p^*	0.303	0.932	0.766	0.265
	N	3,430	2,930	1,441	1,232

Table $B1$: Deterministic tests of $H_0 : p_L^i = p_R^i \ \forall i$ for various subsamples, $B = 1000$

For male players, the deterministic test reaches the same conclusions as the randomized Fisher exact test: the joint null hypothesis that winning probabilities are equalized is not rejected for either the whole sample or any of the subsamples. For female players, the deterministic test rejects the joint null for second serves, but it does not reject the null for non-top receivers, as did the KS test based on the Fisher exact test t-values. All these results are consistent with the smaller statistical power of the deterministic test, which we establish in the next section, compared to the KS test based on the randomized t-values.

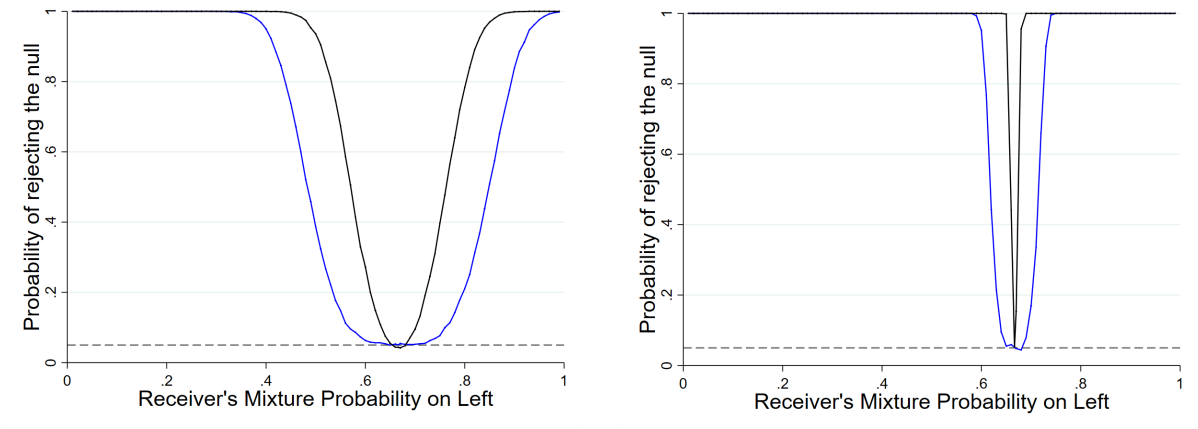
The Power of the Test

We now study the power of the deterministic test, performing for this test the same simulations reported in Section 6 for the KS test based on the randomized Fisher exact test t-values. Figure B1 is the analog of Figure 14 and the power functions in black (the upper curves) reproduce the power functions for the KS test based on the randomized Fisher exact test t-values.

The power functions for the deterministic test, shown in blue, are generated as follows. Let $\theta \in [0, 1]$. We first simulate a random sample under the alternative hy-

pothesis $H_a(\theta) : p_L^i = p_L(\theta)$ and $p_R^i = p_R(\theta)\ \forall i \in \{1, \ldots, N\}$. Let $\hat{X} = \sum_{i=1}^{N} |\hat{p}_L^i - \hat{p}_R^i|$ be the associated value of the test statistic, and let n_S, n_L, and n_R be the associated vectors of marginal distributions. We then simulate 1000 values of the test statistic under the null hypothesis that $p_L^i = p_R^i\ \forall i \in \{1, \ldots, N\}$. Let $\hat{F}^*_{|n_S,n_L,n_R}(X)$ be the empirical *c.d.f.* of the simulated test statistic. The null hypothesis is rejected at the 5% significance level if either $\hat{F}^*_{|n_S,n_L,n_R}(\hat{X}) \leq .025$ or $1 - \hat{F}^*_{|n_S,n_L,n_R}(\hat{X}) \leq .025$, i.e., if the realized value of the test statistic for the simulated sample is in the tails of the *c.d.f.* of the simulated test statistic. The power function for the deterministic test is the probability that the null is rejected when $H_a(\theta)$ is true.[30]

It is evident from Figure B1 that the deterministic test has substantially less power, for both small and large samples. These results demonstrate the usefulness of the test we develop in the body of the paper. Given the low power of the deterministic test, unsurprisingly it fails to reject the null for some sub-samples for which the randomized Fisher exact test does reject.



$N = 40$ $N = 7000$

Figure B1: Power functions for the deterministic test and for the KS test based on t-values

Table B2 shows that the deterministic test is especially low powered, in comparison to our KS test based on the randomized Fisher exact test t-values, in the neighborhood

[30] For the power functions reported in Figure B1 and Table B1, when $N = 40$ and when $N = 7000$, the data is simulated 10,000 times and 1000 times, respectively, for each value of $\theta \in \{0, .01, .02, \ldots, .99, 1\}$ and for $\theta = 2/3$.

of the null hypothesis.

True θ	KS based on t's	Deterministic $\hat{X} = \sum_{i=1}^{N} \lvert\hat{p}_L^i - \hat{p}_R^i\rvert$
0.65	0.997	0.052
0.66	0.460	0.051
2/3	0.046	0.057
0.67	0.153	0.056
0.68	0.964	0.043

Table B2: Rejection rate for H_0 at the 5% level, $N = 7000$

10 Appendix C: Robustness of Power Simulations

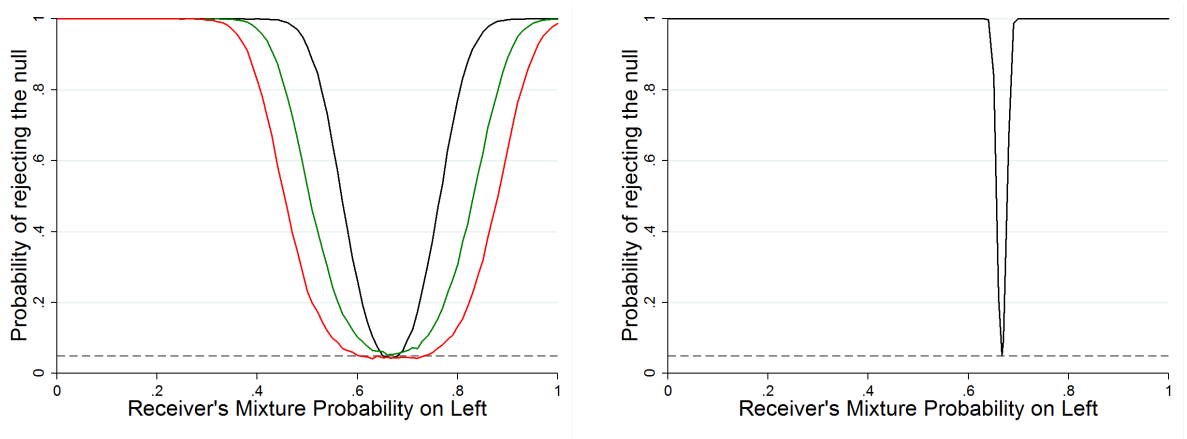
As a robustness check, here we reproduce the simulation results reported in Section 6, but where now the simulated data matches the characteristics of the observed data, point game by point game, rather than just in aggregate. Specifically, if point game i has n_R^i serves to the right, n_L^i serves to the left, and an empirical winning frequency of $\hat{p}^i$, then the simulated data for point game i has n_R^i serves to the right, n_L^i serves to the left, and the probability of winning a point is $\hat{p}^i$ for serves in each direction (and hence the null hypothesis that $p_L^i = p_R^i$ is true). The number of winning serves to the right and left are therefore distributed, respectively, $B(n_R^i, \hat{p}^i)$ and $B(n_L^i, \hat{p}^i)$ in the simulated data for point game i.

The Power of Our Test

The subsection "The Power of Our Test" in Section 6 provided the power functions for the Pearson joint test and the KS tests based on the Pearson p-values and the Fisher exact t-values. It demonstrated that for "small" samples of 40 point games, the test based on the t-values was substantially more powerful than the other two. In addition, for "large" samples of 7000 point games, the test based on the t-values was extraordinarily powerful – the joint null hypothesis of equality of winning probabilities is almost surely rejected for even small departures from equilibrium play.

Figure C1 is the analogue to Figure 14 and shows that the power functions in Figure 14 are largely unchanged when the data is simulated (under the null hypothesis) to match characteristic of the WW data (Figure C1(a)) or the Hawk-Eye data

(Figure C1(b)).



(a) $N = 40$ (b) $N = 7198$

Figure C1: Power Functions for KS test based on t-values (black), p-values (red), and Pearson joint (green)

Table C1 is the analogue Table 6. Comparing to the two tables reveals that the KS test based on the t's is slightly less powerful when the simulated data matches the characteristics of the Hawk-Eye data. This is a consequence of the fact that there were 70 serves per point game for the simulation results reported in Table 6, while there are only 33 serves, on average, per point game in the Hawk-Eye data. As noted previously, the Pearson goodness of fit p-values are only asymptotically uniformly distributed Hence it is unsurprising that the Pearson joint test and the KS test based on the Pearson p-values perform poorly given the smaller number of serves. Table C1 shows that the (true) joint null hypothesis of equality of winning probabilities is rejected, at the 5% significant level, for sure by the KS test based on the p's and it is rejected with probability .726 by the Pearson joint test. These

results reaffirm our conclusion that these tests are not valid for large samples.

True θ	KS based on t's	KS based on p's	Pearson joint test
0.65	0.834	1	0.746
0.66	0.212	1	0.716
2/3	0.051	1	0.726
0.67	0.093	1	0.726
0.68	0.675	1	0.764

Table C1: Rejection rate for H_0 at the 5% level, $N = 7198$

10.1 Ball Bounces

Figure C2 below shows actual and imputed ball bounces for male second serves from the deuce court.

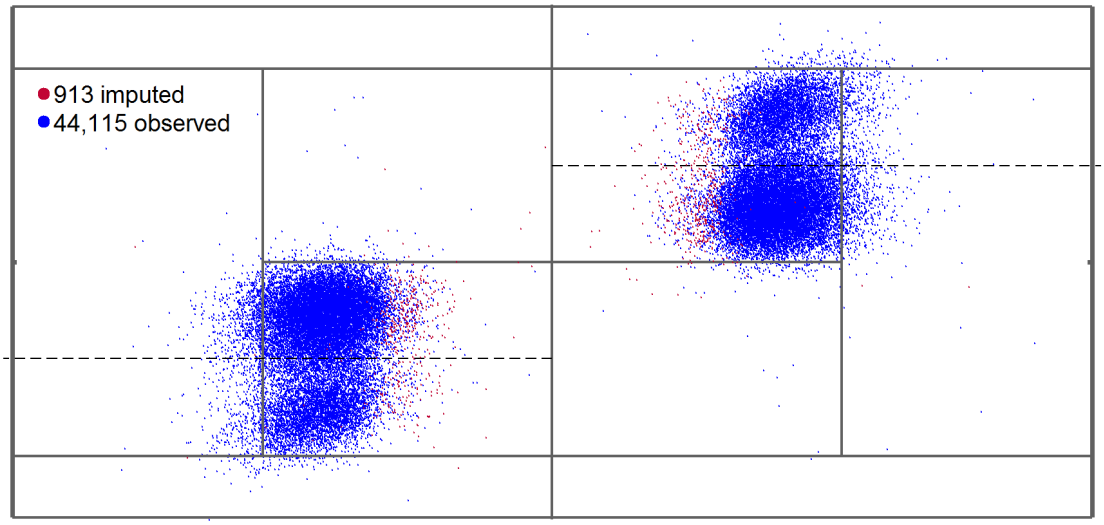


Figure C2: Ball Bounces for Deuce Court Second Serves by Men

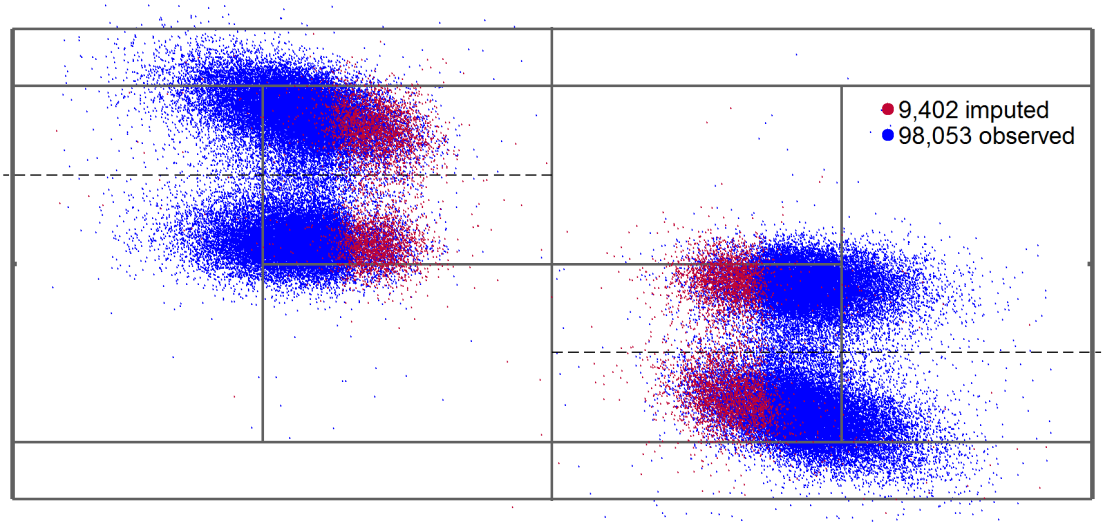


Figure C3: Ball Bounces for Ad Court First Serves by Men

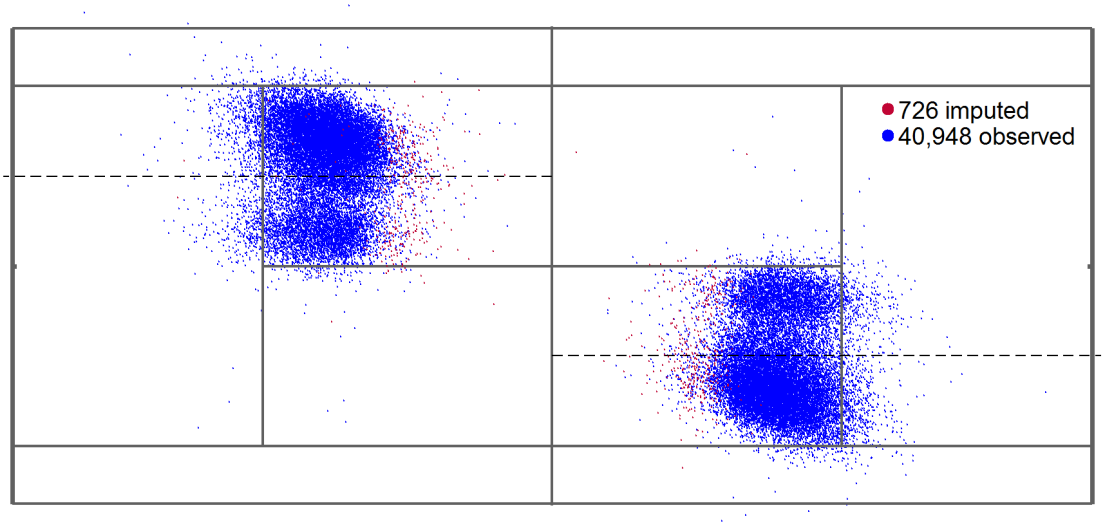


Figure C4: Ball Bounces for Ad Court Second Serves by Men

Ball bounces for first and second serves by women are below.

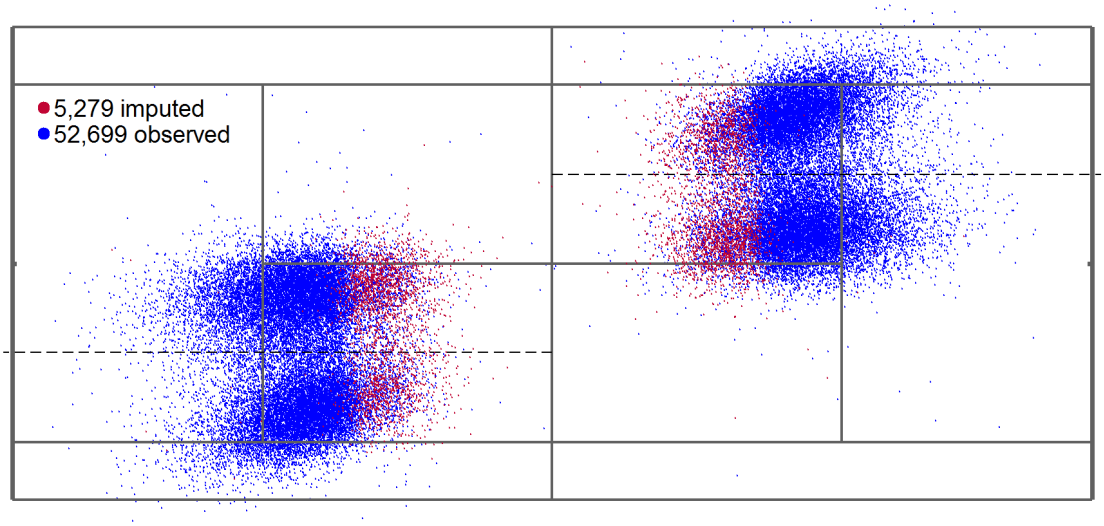


Figure C5: Ball Bounces for Deuce Court First Serves by Women

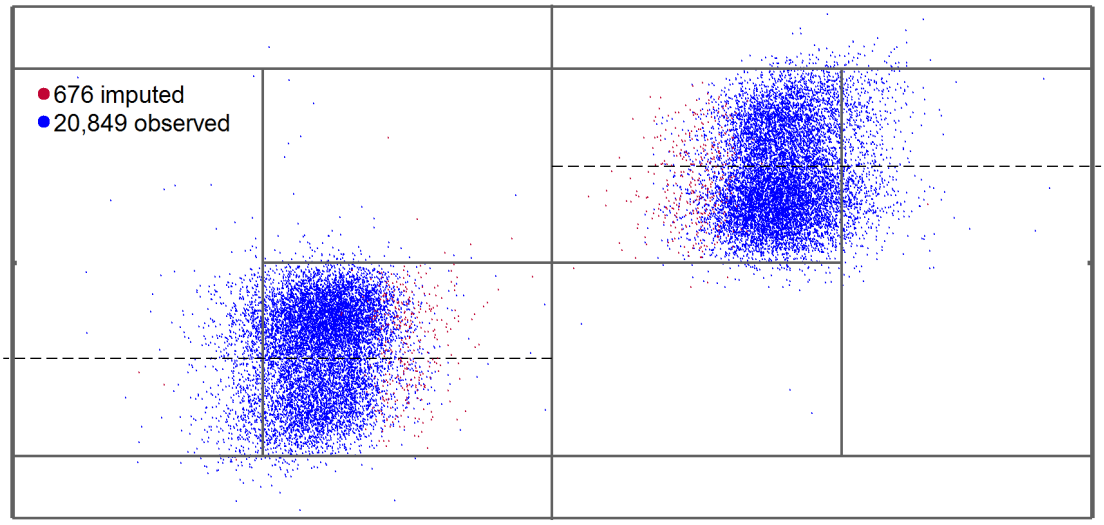


Figure C6: Ball Bounces for Deuce Court Second Serves by Women

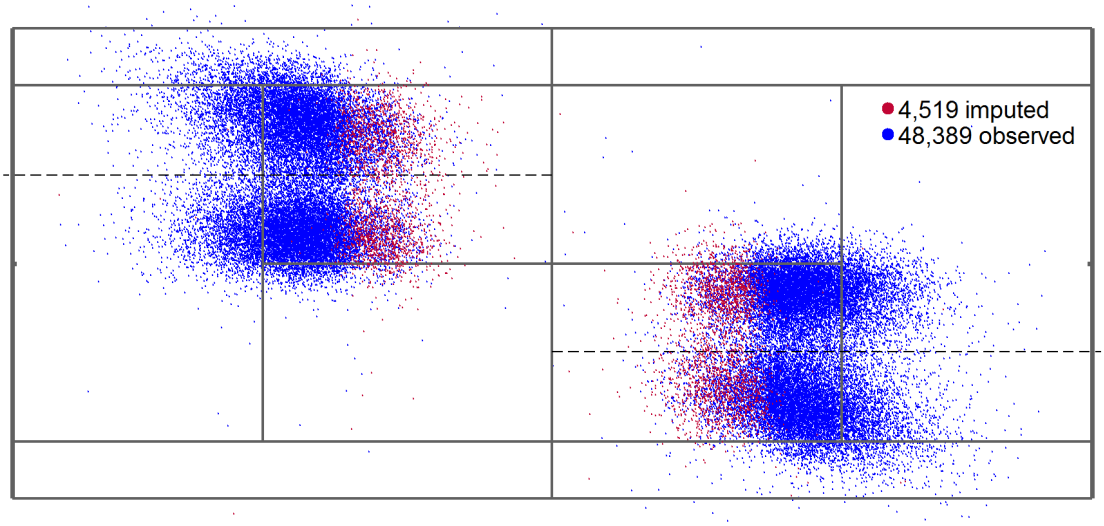


Figure C7: Ball Bounces for Ad Court First Serves by Women

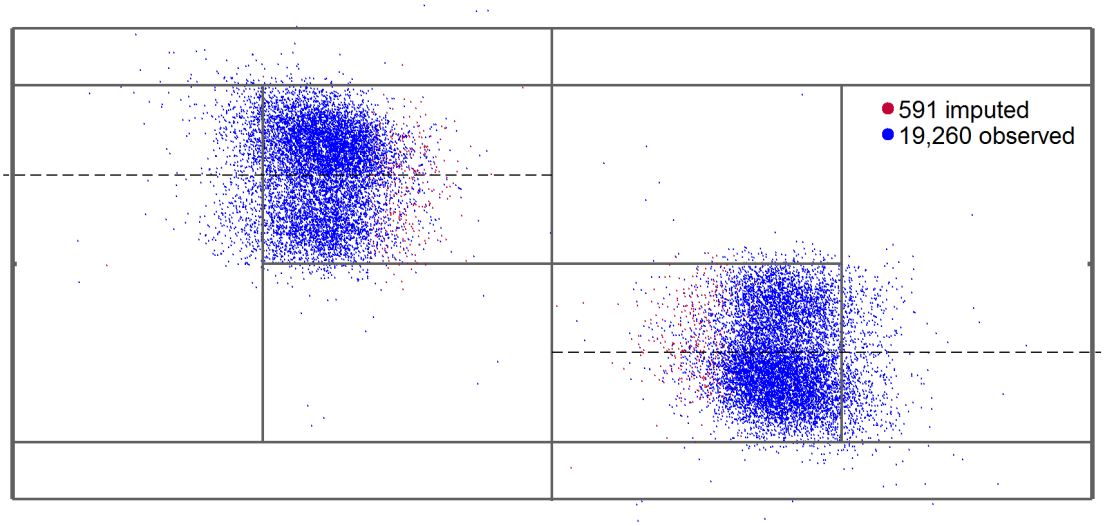


Figure C8: Ball Bounces for Ad Court Second Serves by Women

References

[1] Brown, D. and R. Rosenthal (1990): "Testing the Minimax Hypothesis: A Re-examination of O'Neill's Experiment," *Econometrica* **58**, 1065-1081.

[2] Camerer, C. (2003): Behavioral Game Theory Experiments in Strategic Interaction, Princeton University Press.

[3] Chiappori, P., S. Levitt, and T. Groseclose (2002): "Testing Mixed Strategy Equilibria When Players are Heterogeneous: The Case of Penalty Kicks in Soccer," *American Economic Review* **92**, 1138-1151.

[4] Cooper, D., Kagel, J., Lo, W. and Qin Liang Gu (1999): "Gaming Against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers," *American Economic Review* **89**, 781-804.

[5] Fisher (1935): *The Design of Experiments*, New York: Hafner Publishing Company.

[6] Hsu, S., Huang, C. and C. Tang (2007): "Minimax Play at Wimbledon: Comment," *American Economic Review* **97**, 517-523.

[7] Gibbons, J. and S. Chakraborti (2003): *Nonparametric Statistical Inference*, New York: Marcel Dekker.

[8] Gonzalez-Diaz, J., Gossner, O., and B. Rogers (2102): "Performing best when it matters most: Evidence from professional tennis," *Journal of Economic Behavior & Organization*, **84**, 767-781.

[9] Klaassen, F. and J. Magnus (2001): "Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model," *Journal of the American Statistical Association* **96**, 500–509.

[10] Kocher, M., Lenz, M., and M. Sutter (2012): "Psychological Pressure in Competitive Environments: New Evidence from Randomized Natural Experiments," *Management Science* **58**, 1585-1591.

[11] Kovash, K., and S. Levitt (2009): "Professionals Do Not Play Minimax: Evidence from Major League Baseball and the National Football League," NBER working paper 15347.

[12] Levitt, S., List, J., and D. Reiley (2010): "What Happens in the Field Stays in the Field: Professionals Do Not Play Minimax in Laboratory Experiments," *Econometrica* **78**, 1413-34.

[13] Levitt, S., List, J., and S. Sadoff (2011): "Checkmate: Exploring Backward Induction among Chess Players," *American Economic Review* **101**, 975-990.

[14] Lehmann, E. and J. Romano (2005): Testing Statistical Hypotheses, New York: Springer.

[15] MacKinnon, J. (2009): "Bootstrap Hypothesis Testing," 183-210, in Handbook of Computational Econometrics, edited by David Belsey and Erricos John Kontoghiorghes, John Wiley & Sons.

[16] Mood, A., Graybill, F., and D. Boes (1974): *Introduction to the Theory of Statistics*, New York: McGraw Hill.

[17] O'Neill, B. (1987): "Nonmetric Test of the Minimax Theory of Two-Person Zero-Sum Games," *Proceedings of the National Academy of Sciences* **84**, 2106-2109.

[18] O'Neill, B. (1991): "Comments on Brown and Rosenthal's Reexamination," *Econometrica* **59**, 503-507.

[19] Palacios-Huerta, I. (2003): "Professionals Play Minimax," *Review of Economic Studies* **70**, 395-415.

[20] Palacios-Huerta, I. and O. Volij (2008): "Experientia Docent: Professionals Play Minimax in Laboratory Experiments," *Econometrica* **76**, 71-115.

[21] Paserman, D. (2010): "Gender Differences in Performance in Competitive Environments: Evidence from Professional Tennis Players," mimeo.

[22] Rapoport, A. and R. Boebel (1992): "Mixed Strategies in Strictly Competitive Games: A Further Test of the Minimax Hypothesis," *Games and Economic Behavior* **4**, 261-283.

[23] Rapoport, A., Erev, I., Abraham, E., and D. Olsen (1997): "Randomization and Adaptive Learning in a Simplified Poker Game," *Organizational Behavior and Human Decision Processes* **69**, 31-49.

[24] Rosenthal, R., J. Shachat, and M. Walker (2003): "Hide and Seek in Arizona," *International Journal of Game Theory* **32**, pp. 273-293.

[25] Siegel, S. and N. Castellan (1988): Nonparametric Statistics for the Behavioral Sciences, New York: McGraw-Hill.

[26] Tocher, K. (1950): "Extension of the Neyman-Pearson Theory of Tests to Discontinuous Variates," *Biometrika* **37**, 130-144.

[27] Shachat, J. (2002): "Mixed Strategy Play and the Minimax Hypothesis," *Journal of Economic Theory* **104**, 189-226.

[28] Van Essen, M., and J. Wooders (2015): "Blind Stealing: Experience and Expertise in a Mixed-Strategy Poker Experiment," Games and Economic Behavior **91**, 186-206.

[29] Walker, M. and J. Wooders (2001): "Minimax Play at Wimbledon," *American Economic Review* **91**, 1521-1538.

[30] Walker, M., Wooders, J., and R. Amir (2011): "Equilibrium Play in Matches: Binary Markov Games," *Games and Economic Behavior* **71**, pp. 487–502.

[31] Wooders, J. and J. Shachat (2001): "On The Irrelevance of Risk Attitudes in Repeated Two-Outcome Games," *Games and Economic Behavior* **34**, 342-363.

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