Nash at Wimbledon: Evidence from Half a Million Serves[*]

Romain Gauriot[†] Lionel Page[‡] John Wooders[§]

September 2016

Abstract

Minimax and its generalization to mixed strategy Nash equilibrium is the cornerstone of our understanding of strategic situations that require decision makers to be unpredictable. Using a dataset of nearly half a million serves from over 3000 matches, we examine whether the behavior of professional tennis players is consistent with the Minimax Hypothesis. The large number of matches in our dataset requires the development of a novel statistical test, which we show is more powerful than the tests used in prior related studies. 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 for comments from seminar participants at the University of New South Wales and conference participants at GAMES 2016, the 2015 World Meetings of the ESA, the 10th ANZWEE meetings, and EBCC 2016. We are also grateful to Mark Walker for useful comments. Wooders is grateful for financial support from the Australian Research Council's *Discovery Projects* funding scheme (project number DP140103566).

[†] School of Economics, University of Sydney, romain.gauriot@sydney.edu.au.

[‡] School of Economics and Finance, Queensland University of Technology, lionel.page@qut.edu.au.

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

1 Introduction

von Neumann's Minimax Theorem for two-player zero-sum games and Nash's generalization to equilibrium in n-player non-zero sum games are the foundations of modern game theory. Nash equilibrium and its extension to decision making in dynamic settings or settings with incomplete information are routinely applied in theoretical models and are the basis of much of our understanding of strategic interaction. Here we test the predictive accuracy of the theory using data from the field.

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.

Data from professional sports, however, has been far more supportive of theory. Using poker as a motivating example, Walker and Wooders (2001) argue that while the rules of a 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, in contrast, players have typically devoted their lives to the game and they have substantial financial incentives.

The present paper re-evaluates the question of whether the behavior of sports professionals conforms to theory using a unique dataset from Hawk-Eye, a computerized ball tracking system employed at Wimbledon and other top championship tennis matches. It make 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 broader test of the theory by analyzing both first and second serves, for both men and women. Finally, combining the Hawk-Eye data with a dataset on player rankings, it shows that even tennis professionals differ in the degree to which their behavior conforms to theory: the behavior in the field of higher ranked players conforms more closely to theory.

A critique of the results of prior studies has been that they have low power to reject the null hypotheses implied by the theory.[1] Walker and Wooders (2001), for example, study a dataset comprised of approximately 3000 serves made in 10 championship

[1] See Kovash and Levitt (2009), for example.

tennis matches. Chiappori, Levitt, and Groseclose (2002) and Palacios-Huerta (2003) study 459 and 1417 penalty kicks, respectively. Our dataset contains the precise trajectory and bounce points of the tennis ball for nearly 500,000 serves from over 3000 professional tennis matches, and thereby allows for an extremely powerful test of the theory.

The large number of matches in our dataset requires that we develop a novel statistical test for our analysis. When the number of points in each match is small relative to the overall number of matches, as it is in our dataset, we show that the test introduced in Walker and Wooders (2001) is not valid: it rejects the implication of the minimax hypothesis that winning probabilities are equalized, even when the null is true. The new test that we develop, based on the Fisher exact test, rejects the true null hypothesis with probability of exactly 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 Walker and Wooders (2001) 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 valid to run the test once and reject the null hypothesis if the p-value is less than the 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 $x\%$ weight on p-values below .05."

We find that the win rates of male professional tennis players are consistent with the minimax hypothesis. 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 empirical density of p-values places no 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 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 44.9% weight on p-values

[2] The Walker and Wooders test is valid for their data set, in which the number of points in each match is large relative to the number of matches.

below .05 for first serves, and a 16.0% 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 subjects in comparable laboratory tests of the minimax hypothesis. 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 women, despite having vastly more data.

A second implication of the minimax hypothesis is that the players' choices of direction of serve are serially independent. We reject serial independence for both men and women, for both first and second serves. Players switch the direction of their serve too frequently to be consistent with randomness. Negative serial correlation in the direction of serve is more pronounced for women than men, and the difference is statistically significant.

We conjecture that men's greater physical strength 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.[3] 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. It is evident that the serve is relatively more important than in men's than women's tennis. A receiver in men's tennis who fails to play minimax (and equalize the server's winning probabilities) is much 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. Likewise, in men's tennis, a server who is predictable because he fails to randomize in the direction of his serve is much more vulnerable to exploitation by the receiver.

We also find evidence that the behavior of higher ranked (i.e., better) players conforms more closely to the minimax hypothesis. Higher-ranked male players exhibit less serial correlation in their first serve than lower ranked players. For female players, by contrast, rank does not have a statistically significant effect on the degree of serial correlation, again perhaps a consequence of the smaller importance of the serve in

[3] For first serves by men, on average only 0.45 seconds elapses between the serve and the first bounce.

women's tennis. In each case, as one might expect, the rank of the receiver has no statistically significant effect on the degree of serial correlation exhibited by the server.

To further investigate the effect of ability on behavior, we divide the data into two subsamples based on the receiver's rank. (It is important to keep in mind that it is the *receiver's* strategy that determines whether winning probabilities are equalized across directions of serve.) In one subsample the receiver was a "top" player, i.e., above the median rank, and in the other the receiver was 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 the minimax hypothesis on each subsample.

As noted above, win rates conform to the minimax hypothesis somewhat less neatly for women than for men. 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 the minimax hypothesis, while the equality of winning probabilities is resoundingly rejected for the subsample in which the receiver is not a top player. This result show that behavior of women conforms more closely to the minimax hypothesis for better receivers.

Related Literature

Walker and Wooders (2001), henceforth 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, switch the direction of their serve too often to be consistent with the random play predicted by the theory. WW find, however, that professionals deviate far less from random play than do student subjects in comparable laboratory experiments.

Hsu, Huang, and Tang (2007), henceforth HHT, broaden the analysis of WW by considering data from women's and junior's matches in addition to data from men's matches. In a sample of 9 women's matches, 8 juniors matches, and 10 men's matches, HHT find also found that win rates conformed to the theory. 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 18.5% and 49.6%, 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 the minimax hypothesis. In both cases, the empirical density of p-value assigns zero probability to p-values below .05.

Chiappori, Levitt, and Groseclose (2002), henceforth CLG, study a dataset of every penalty kick occurring in French and Italian elite leagues over a three year period (459 penalty kicks), and test whether play conforms to the mixed strategy Nash equilibrium (and minimax solution) of a parametric model of a penalty kick in soccer 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.[4] A key contribution of CLG is the precise identification of the predictions of the minimax hypothesis that are robust to aggregation across heterogeneous contests. It finds that the data conforms to the qualitative predictions of the model, e.g., kickers choose "center" more frequently than goalies.[5]

Palacios-Huerta (2003) studies a group of 22 kickers and 20 goalkeepers who have participated in a relatively large number of penalties (each 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.[6] Importantly, his analysis ignores that a kicker generally faces different goalkeepers (and different 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

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

[5] 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.

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

a period of several hours.[7] It is plausible therefore to assume that 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 the minimax hypothesis. There is no need to aggregate data as in CLG or PH.

The present paper is less closely 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), Van Essen and Wooders (2015)). Palacios-Huerta and Volij (2010) report evidence that professional soccer players behave according to the minimax hypothesis 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 (2010)'s own data is inconsistent with the minimax hypothesis.

In Section 2 we present the model of a serve in tennis and the testable hypotheses of the minimax hypothesis. In Section 3 we describe our data. In Section 4 we describe our test of the hypothesis that winning probabilities are equalized and we present our results, while in Section 5 we report the results of our test that the direction of serve is serially independent. In Section 6 we establish that the behavior of higher ranked players conforms more closely to theory than for lower ranked players. In Section 7 we show that (i) the WW test of equality of winning probabilities 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 when 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 the test to the data from HHT.

2 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 (L) or right (R). 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 as in Figure 1.

[7] 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.

		Receiver	
		L	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. The probability payoffs differ for men and women.[8] At the first serve, the probability payoffs include the possibility that the server ultimately wins the point after an additional (second) serve.

If the first serve is a fault, then the server gets a second, and final, serve. The server chooses whether to serve L or R and the receiver simultaneously chooses whether to overplay L or R. If the second serve is also a fault, then server loses the point. 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]

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

[8] As noted in the Introduction, men win 64% of the points when they have the serve, while women win only 56%.

[9] Indeed, first and second serves are played differently. In our data set, the average speed of a first serve is 160 kph and of the second serve is 126 kph (35.3% of first serves fault, but only 7.5% of second serves fault).

serve. Under this assumption there is a unique Nash equilibrium and it is in (strictly) mixed strategies.[10]

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 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 minimax (and equilibrium) 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}. 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 any other aspect of the history of play prior to that point.

Two testable implications come from the theory. According to the minimax hypothesis, a player obtains the same payoff from all actions chosen with positive probability. Thus, the server's payoff at the first serve, i.e., the probability of winning the point at the first serve, is the same for serves left and for serves right, when delivered from the same court. Likewise, his probability of winning the point at the second serve is the same for serves left and serves right, when delivered from the same court. A second implication of the theory, which comes from the analysis of the extensive form game representing a match, is that the direction of the serve is serially independent.

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.

[10] Nash equilibrium and minimax coincide in two-player constant sum games, such as this one.

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

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.

[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.

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 2 is an 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 2, such 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 2) 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 2) is classified as a left serve, while a bounce above is classified as a right serve.[13]

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.

[13] More precisely, Hawk-Eye records each impact of the ball with the court by 50 coordinate pairs (x, y) that describe the perimeter of the elipse-shaped bounce point, where all distances are measured in meters from the center of the court (with coordinates $x = y = 0$). A tennis court is 27 feet (8.2296 meters) wide. The deuce court is half as wide, i.e., 4.1148 meters. Hence in the deuce court on the left hand side of Figure 2 we have $-4.1148 \leq y \leq 0$. To classify the direction of the serve, we select one of the 50 (x, y) pairs at random. For $x < 0$, a serve is to the right if $y > -2.0574$ and to the left if $y < -2.0574$.

same theoretical winning probability.

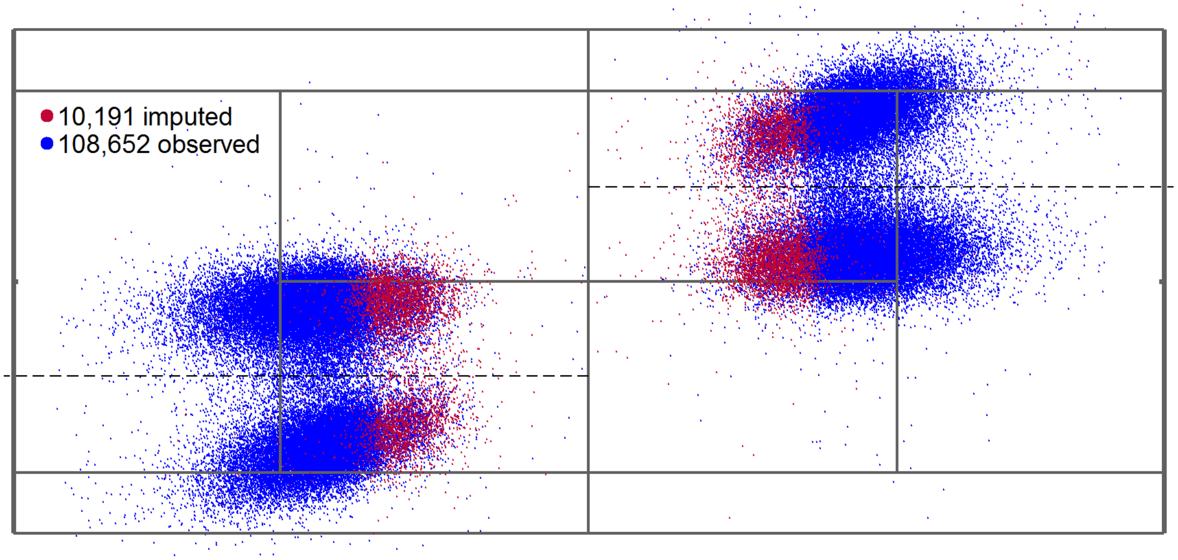


Figure 2: Ball Bounces for Deuce Court First Serves by Men

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 B, Figure B5.

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

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 the minimax hypothesis, the probability that the server wins the point is the same for serves left and for serves right. In the data, for each "point game" i we observe the number of serves to the left and right, n^i_L and n^i_R. Let n^i_{jS} and n^i_{jF} denote the number of serves in direction $j \in \{L, R\}$ for which the ultimate outcome was S (success) – the server won the point, or F (failure) – the receiver won the point, respectively. Thus the data for point game i can be represented in the table below

$$\begin{array}{c|cc|l} & L & R & \\ \hline S & n^i_{LS} & n^i_{RS} & n^i_S = n^i_{LS} + n^i_{RS} \\ F & n^i_{LF} & n^i_{RF} & n^i_F = n^i_{LF} + n^i_{RF} \\ \hline & n^i_L = n^i_{LS} + n^i_{LF} & n^i_R = n^i_{RS} + n^i_{RF} & \end{array}$$

We first describe our test for whether the server's winning probabilities are equal in an individual point game. From this test we construct a test of that winning probabilities are the same for serves right and serves left in every point game.

Individual Play and The Fisher Exact Test

Let p^i_j denote the true, but unknown, probability that the server will win the point when the first serve is in direction j. We use the Fisher exact test to test the null hypothesis that $p^i_L = p^i_R = p^i$ 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. The beauty of the Fisher exact test is that it does not require knowledge of the true (but unknown) value of p^i. Moreover, it is an exact test as does not rely on the asymptotic distribution of the test statistic. As we shall see, the later is essential for constructing a valid test of the joint hypothesis of the equality of winning probabilities, given the large number of point games in our sample.

Let $f(n_{LS}; n_S, n_L, n_R)$ denote the probability, under the null, that the server wins n_{LS} serves to the left, conditional on winning n_S serves in total, after delivering n_L and n_R serves to the left and to the right, respectively, i.e., $f(n_{LS}; n_S, n_L, n_R) \equiv \Pr(n_{LS}|n_S, n_L, n_R)$. This conditional probability is computed as follows:

$$f(n_{LS}; n_S, n_L, n_R) \equiv \frac{\Pr(n_{LS}, n_S, n_L, n_R)}{\Pr(n_S, n_L, n_R)} = \frac{B(n_{LS}; n_L, p)B(n_{RS}; n_R, p)}{B(n_S; n_L + n_R, p)},$$

where $n_{RS} = n_S - n_{LS}$ and $B(n_{jS}; n_j, p)$ is the binomial probability of winning n_{jS} of n_j serves in direction $j \in \{L, r\}$ when the winning probability is p. The equality follows from the fact that the binomial processes for serves left and serves right are independent. By direct calculation we have

$$\begin{aligned} f(n_{LS}; n_S, n_L, n_R) &= \frac{\binom{n_L}{n_{LS}} p^{n_{LS}} (1-p)^{n_L - n_{LS}} \binom{n_R}{n_{RS}} p^{n_{RS}} (1-p)^{n_R - n_{RS}}}{\binom{n_L + n_R}{n_S} p^{n_S} (1-p)^{n_L + n_R - n_S}} \\ &= \frac{\binom{n_L}{n_{LS}} \binom{n_R}{n_{RS}}}{\binom{n_L + n_R}{n_S}}. \end{aligned}$$

Of critical importance, this conditional probability is exact for finite samples and does not depend on p.[14] Let $F(n_{LS}; n_S, n_L, n_R)$ be the associated *c.d.f.*, i.e.,

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

In its standard application, the Fisher exact test rejects the null hypothesis at the 5% significance level if $F(n_{LS}; n_S, n_L, n_R) \leq .025$ or $F(n_{LS}; n_S, n_L, n_R) \geq .975$. For any given marginal distribution (identified by n_S, n_L, and n_R), since the density f is discrete, a 5% test will not typically have a size of exactly 5%.

We employ a randomized test in order to obtain a test of exactly the correct size. For each point game i, let t^i be the random test statistic given by a draw from the distribution $U[0, F(n^i_{LS}; n^i_S, n^i_L, n^i_R)]$ if n^i_{LS} takes its minimum value, i.e., $n^i_{LS} = n^i_S - n^i_R$, and from the distribution $U[F(n^i_{LS} - 1; n^i_S, n^i_L, n^i_R), F(n^i_{LS}; n^i_S, n^i_L, n^i_R)]$

[14] The density is defined for $n_{LS} \in \{\max(n_S - n_R, 0), \ldots, \min(n_S, n_L)\}$. We require in particular that $n_{LS} \geq \max(n_S - n_R, 0)$, i.e., the number of winning left serves must be (i) non-negative, and (ii) at least as great at the total number of winning serves minus the number of right serves. Likewise, we require $n_{LS} \leq \min(n_S, n_L)$, i.e., the number of winning left serves can not exceed either the number of winning serves overall or the number of left serves.

otherwise. Under the null hypothesis that $p_L^i = p_R^i$, the test statistic t^i is distributed $U[0,1]$.[15] Hence rejecting the null hypothesis if $t^i \leq .025$ or $t^i \geq .975$ yields a test of exactly size .05. We refer to this test as the randomized Fisher exact test. The t^i's obtained from this test will be used in the next section to test the joint null hypothesis that $p_L^i = p_R^i$ for every i.

An Illustrative Example

Consider the hypothetical data below, for three different point games, all of which have the same marginal distributions.

$$\begin{array}{c|cc|c} & L & R & \\ \hline S & \mathbf{4} & 6 & 10 \\ F & 11 & 0 & 11 \\ \hline & 15 & 6 & \end{array} \qquad \begin{array}{c|cc|c} & L & R & \\ \hline S & \mathbf{8} & 2 & 10 \\ F & 7 & 5 & 11 \\ \hline & 15 & 6 & \end{array} \qquad \begin{array}{c|cc|c} & L & R & \\ \hline S & \mathbf{10} & 0 & 10 \\ F & 5 & 6 & 11 \\ \hline & 15 & 6 & \end{array}$$

$$\begin{array}{lll} f(4;10,15,6) = .0039 & f(8;10,15,6) = .2737 & f(10;10,15,6) = .0085 \\ F(4;10,15,6) = .0039 & F(8;10,15,6) = .9063 & F(10;10,15,6) = 1.00 \end{array}$$

The hypothetical data considers three ($n_{LS} = 4$, 8, and 10) of 7 possible realizations of n_{LS} (i.e., $n_{LS} = 4, \ldots, 10$) consistent with the marginal distributions of the example.

For this hypothetical data, at the 5% significance level the Fisher exact test rejects the null hypothesis that $p_L^i = p_R^i$ for the data only for the left-most ($n_S = 4$) and the right-most ($n_S = 10$) tables: the data in the left-most table favors the alternative hypothesis that $p_L^i < p_R^i$, while the data on the right-most table favors the alternative $p_L^i > p_R^i$. The size of this test, however, is only $.0039 + .0085 = 0.0124$.

The randomized Fisher exact test likewise rejects the null hypothesis with probability 1 if $n_S = 4$ or $n_S = 10$. If $n_S = 5$ then t^i is a drawn from

$$U[F(4;10,15,6), F(5;10,15,6)] = U[.0039, .0550]$$

and the null is rejected if $t^i \leq .025$. This occurs with probability .413. Likewise, if $n_S = 9$ then $t^i \backsim U[.9064, .9915]$ and the null is rejected if $t^i \geq .975$. This occurs with probability .194. Otherwise, for $n_S \in \{6, 7, 8\}$, we have $.025 < t^i < .975$ and the null is not rejected. Hence the size of the test is

$$f(4;10,15,6) + .413 f(5;10,15,6) + .194 f(9;10,15,6) + f(10;10,15,6) = .05.$$

[15] The proof is elementary and omitted. See Wooders (2008) footnote 9 for the proof of the analogous result for the randomized binomial test.

One can think of a realization of $n_S = 5$ or $n_S = 9$, while not being sufficiently extreme to lead to an outright rejection the null, as providing some evidence against it and thus the null is rejected with positive probability.

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. 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 5000 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 $H_0 : p^i_L = p^i_R$ (5000 trials)

At the individual level, the rate at which equality of winning probabilities is rejected at the 5% or 10% significance level seems – to the eye – to be roughly consistent with the theory.

Aggregate Play and the Joint Null Hypothesis

We next consider the joint null hypothesis that $p^i_L = p^i_R$ for each point game i. WW test this hypothesis by applying the Kolmogorov–Smirnov (KS) test to the empirical distribution of the p-values, one for each point game, obtained from the

[16] Since each point game has fewer observations of 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.

Pearson Goodness of fit test. Under the joint null hypothesis, each p-value is asymptotically uniformly distributed. They find that the null is resoundingly rejected for O'Neill's (1987) experimental data, whereas the null is not rejected when applied to the serve and return data from professional tennis.[17] Other authors have followed this approach to test for equality of winning probabilities in professional soccer (Palacios-Huerta (2003)) and laboratory experiments with human subjects (Levitt-List-Reiley (2010), Van Essen and Wooders (2015)).

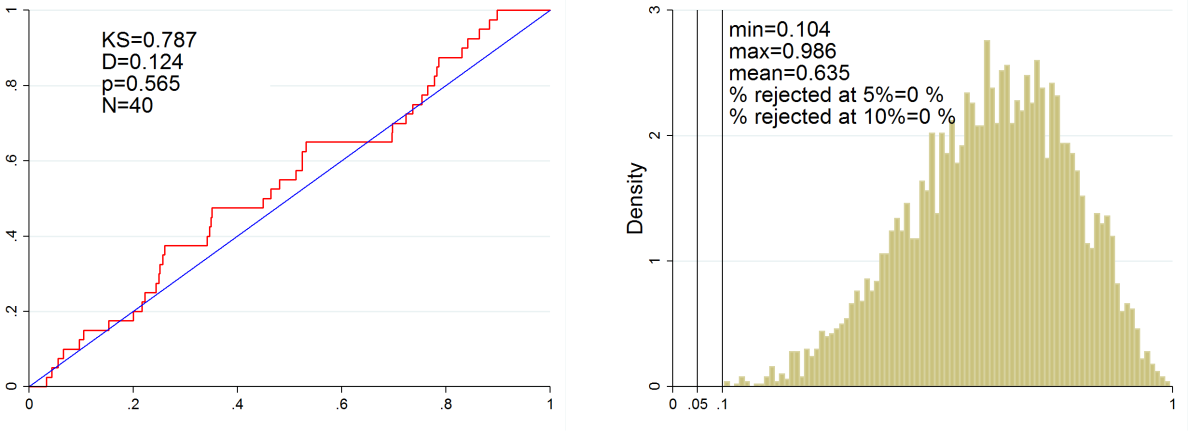
Here we construct a new test of the joint null hypothesis that $p_L^i = p_R^i$ for each point game i. Since our test is constructed from the randomized Fisher exact test, it has the crucial advantage that the t values (the analogue to the p-values in WW) are uniformly distributed under the null hypothesis for finite samples. In particular, it does not rely on the asymptotic distribution of a test statistic. Moreover, as we will show later, the test we construct is more powerful than the WW test.

To illustrate, we begin by applying this new test to the WW data. As described earlier, for each point game i, let t^i be the random test statistic given by a draw from the distribution $U[0, F(n_{LS}^i; n_S^i, n_L^i, n_R^i)]$ if n_{LS}^i takes its minimum value of $n_{LS}^i = n_S^i - n_R^i$, and 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 $U[0,1]$. Since the t^i's are *i.i.d.* draws from a continuous distribution, we can test the joint hypothesis by applying the 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 the 40 t-values (one for each point game in the WW dataset), denoted $\hat{F}(x)$, is given by $\hat{F}(x) = \frac{1}{40}\sum_{i=1}^{40} I_{[0,x]}(t^i)$, 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 $K = \sqrt{40}\sup_{x\in[0,1]} |\hat{F}(x) - x|$ has a known asymptotic distribution (see p. 509 of Mood, Boes, and Graybill (1974)).

Figure 3(a) shows one realization of the empirical distribution of t values for the WW data. For this realization, the value of the test statistic is $K = .787$ and the associated p-value is .565. Thus the new test confirms the WW finding that the joint null hypothesis of equality of winning probabilities for first serves does not even come

[17] For the 50 "point games" in O'Neill's data, the KS test statistic is $k = 1.704$ (p-value of .006), while for the 40 point games in Walker and Wooders' data set the KS test statistic is $k = .670$ (p-value of .76).

close to being rejected for male professional tennis players.[18]



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

(b) Density of KS test p-values (5000 trials)

Figure 3: KS test of $H_0 : p^i_L = p^i_R \; \forall i$ (WW data)

Since the t values are stochastic, the empirical *c.d.f.* and the KS test p-value reported in Figure 3(a) are also random. It is natural to question the robustness of the conclusion that the joint null hypothesis is not rejected. Figure 3(b) shows the empirical density of the p-values obtained after 5000 repetitions of the KS 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 5000 p-values were equally distributed across bins, then there would be 50 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 50. By construction, the area of the shaded region in Figure 3(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 3(b) shows that the p-values are concentrated around .6, and hence are far from the rejection region. In 5000 repetitions of the KS test, the joint null hypothesis of equality of winning probabilities is not once rejected at the 10% significance level. Hence the failure to reject the joint null hypothesis of equality of winning probabilities is indeed completely robust to the realization of the t values.

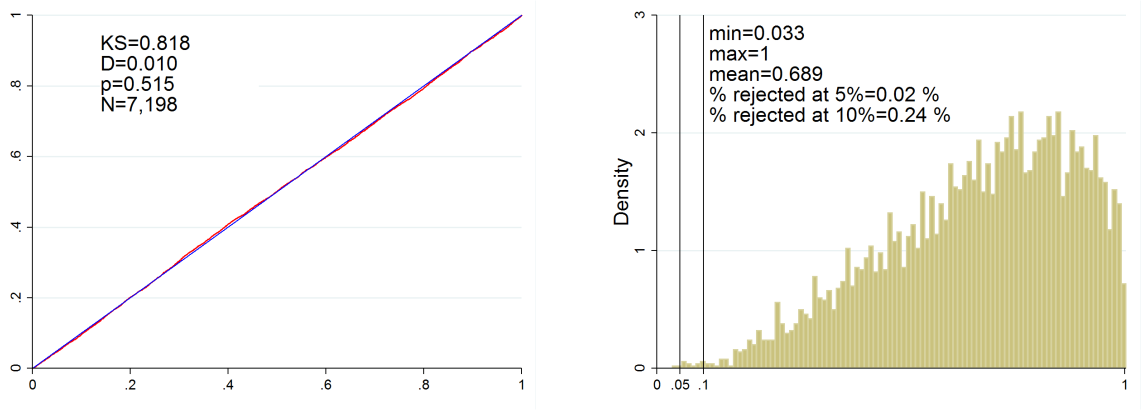
Before proceeding it is important to emphasize several aspects of our test. First, it is a valid test in the sense that if the null hypothesis is true (i.e., $p^i_L = p^i_R \; \forall i$) then the p-value obtained from the KS test is asymptotically uniformly distributed

[18] In WW (2001), the value of the KS test statistic was .670 and the associated p-value was .76.

(as the number of point games grows large). Second, once the data has been realized, the distribution of p-values obtained from running the test repeatedly (as in Figure 3(b)) depends on the data, and there is no longer reason to expect the KS p-values to be uniformly distributed. Finally, as the number of serves in each point game grows large, then the intervals $U[F(n^i_{LS}-1;n^i_S,n^i_L,n^i_R),F(n^i_{LS};n^i_S,n^i_L,n^i_R)]$ from which the t values are drawn shrink and the empirical density of the KS p-values collapses to a degenerate distribution.

Figure 4 shows the result of applying the same test to the Hawk-Eye data for first serves by men. As noted in Table 2, this test is based on 226,298 first serves in 7,198 point games. Figure 4(a) shows a realization of the empirical *c.d.f.* of t values (in red) and the theoretical *c.d.f.* (in blue). 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. Despite its enormous power, the test does not come close to rejecting the null hypothesis.

Figure 4(b) shows that this conclusion is robust to the realizations of the t values. In only one instance (.02%) of 5000 trials of the KS test is the null hypothesis rejected at the 5% level. In only .24% of the trials is it rejected at the 10% level. The mean p-value is .689, which is far from the rejection region.

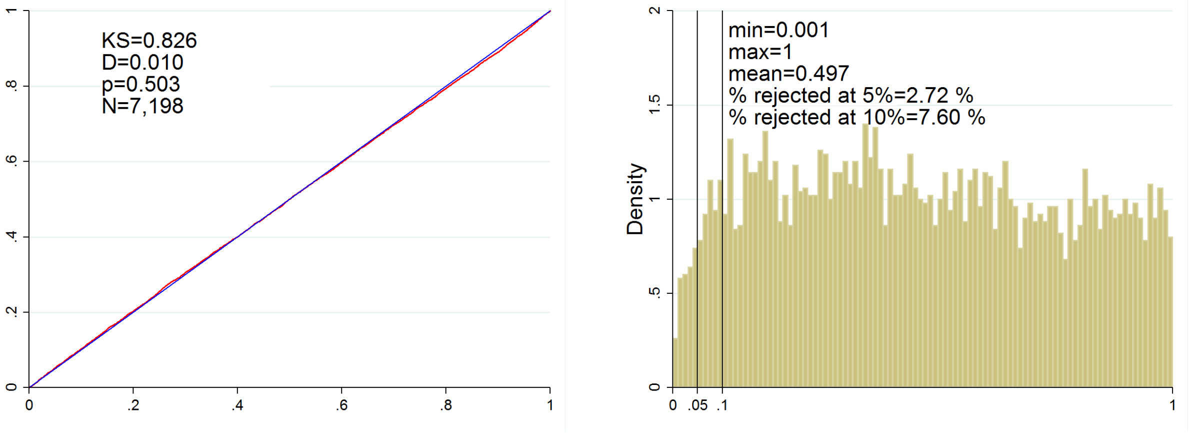


(a) An empirical *c.d.f.* of t-values

(b) Density of KS test p-values (5000 trials)

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

WW studied only first serves since there were few second serves in their data. 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 p-values from the KS tests after 5000 trials. Only for a small fraction of these trials (2.72%) is the joint null rejected at the 5% level. The mean p-value is .497.



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

(b) Density of KS test p-values (5000 trials)

Figure 5: KS test for Men of $H_0 : p^i_L = p^i_R \; \forall i$ (Hawk-Eye, Second Serves)

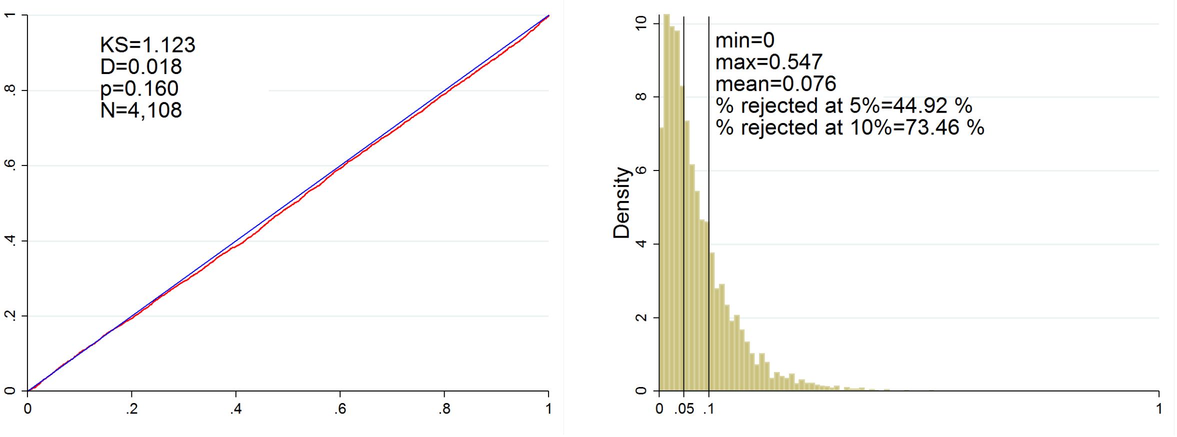
While the data for both first and second serves is strikingly consistent with the theory, comparing Figures 4(b) and 5(b) reveals that the result is slightly less robust for second serves. This is a consequence of the fact that there are fewer second serves than first serves in each point game. Thus the intervals $U[F(n^i_{LS} - 1; n^i_S, n^i_L, n^i_R), F(n^i_{LS}; n^i_S, n^i_L, n^i_R)]$ 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 random.[19]

Our data also allow a powerful test of whether the play of women conforms to the minimax hypothesis. 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 $p^i_L = p^i_R$ for all i, for first and second serves, respectively. For first serves, the null is rejected at the 5% and 10% significance level in 44.92% and 73.46% of 5000 trials, respectively.

The results for second serves are more ambiguous. While the p-values shown in

[19] In a point game with n_L left serves, n_R right serves, and n_S successful serves, there are $\min(n_S, n_L) -$ $\max(n_S - n_R, 0) + 1$ distinct intervals from which t values are drawn.

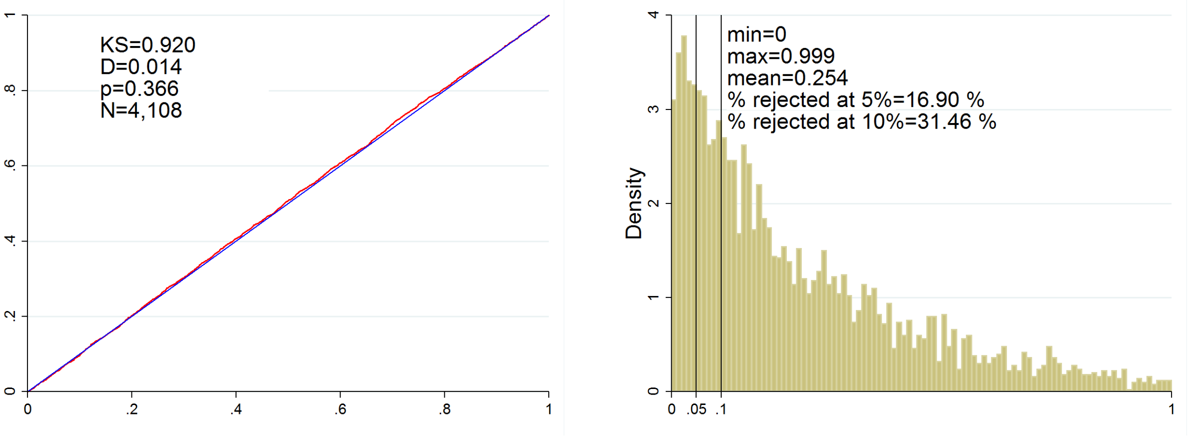
Figure 7(b) tend to be small, the mean p-value is .254. The null hypothesis tends not to be rejected at the 5% level: in only 16.90% of the trials is the p-value below .05.



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

(b) Density KS test p-values (5000 trials)

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



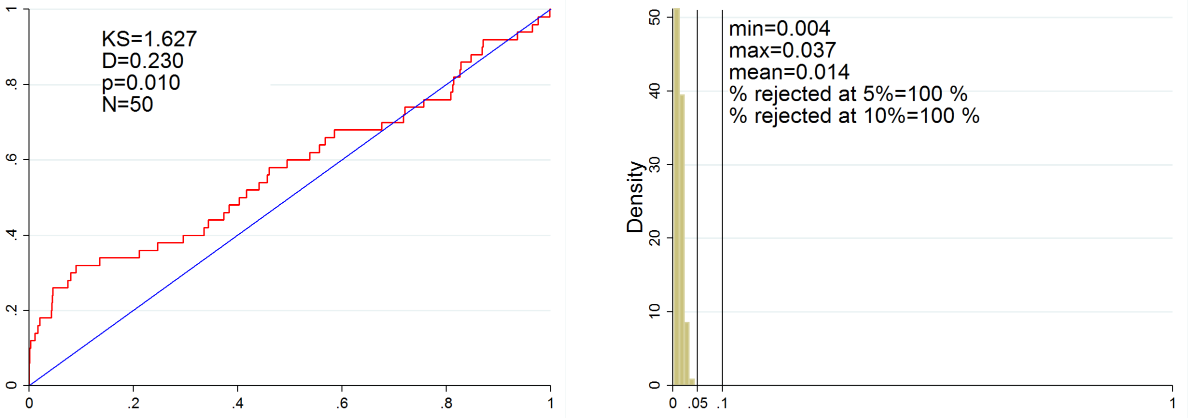
(a) An empirical *c.d.f.* of t-values

(b) Density of KS test p-values (5000 trials)

Figure 7: KS test for Women of $H_0 : p^i_L = p^i_R \ \forall i$ (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. 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 (5000 trials)

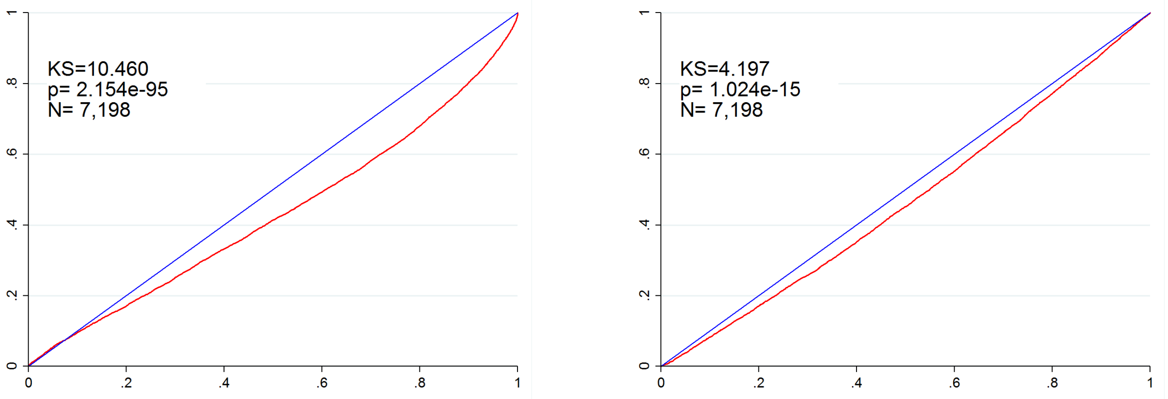
Figure 8: KS test of $H_0 : p^i_L = p^i_R \ \forall i$ (O'Neill's (1987) experimental data)

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_L occurrences of L and n_R occurrences of R is known. Denote this probability by $f_R(r; n_L, n_R)$ and let $F_R(r; n_L, n_R)$ denote the associated *c.d.f.* At the 5% significance level, the null is rejected if $F_R(r; n_L, n_R) \leq .025$ or if $1 - F_R(r - 1; n_L, n_R) \leq .025$, 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, we employ the randomized test introduced in WW. In particular, for each point game i we draw the random test statistic t^i given by a draw from $U[F_R(r^i-1; n^i_L, n^i_R), F_R(r^i; n^i_L, n^i_R)]$. 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 9 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.[20] (We omit the empirical density of the KS test p-values since the null is rejected for every realization of the t's.) 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 much switching, i.e., there are more than the expected number of large t-values. These results confirm the WW finding of negative serial correlation of first serves by men.



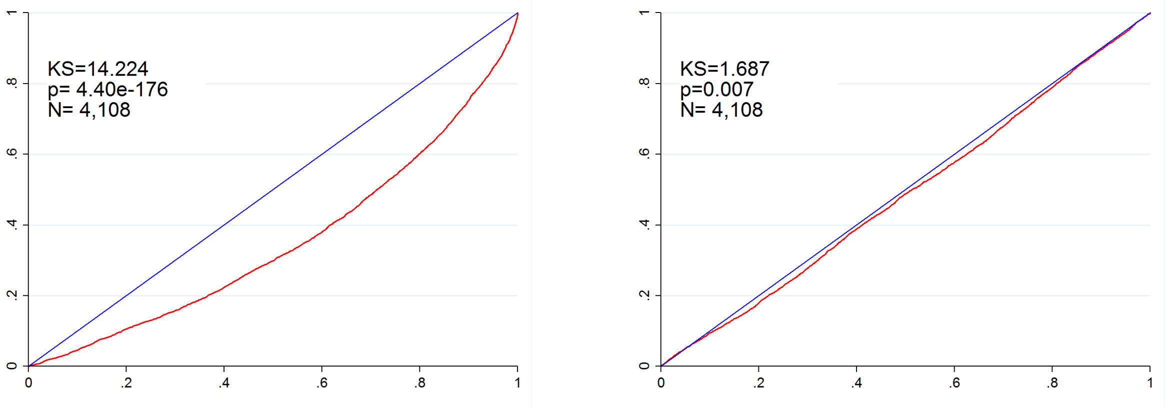
(a) First Serve: Empirical *c.d.f.* of t-values (b) Second Serve: Empirical *c.d.f.* of t-values

Figure 9: KS test for Men of $H_0 : s^i$ is serial independent $\forall i$ (Hawk-Eye)

Figure 10 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 correla-

[20] 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.

tion 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 10: KS test for Women of $H_0 : s^i$ is serial independent $\forall i$ (Hawk-Eye)

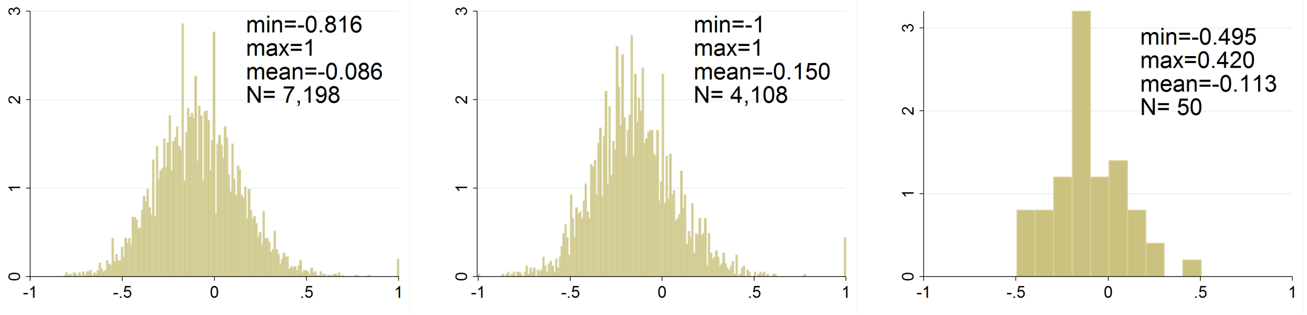
Comparing Figures 10(a) and 11(a) one might be tempted to conclude the women exhibit more serial correlation in first serves than men. While this conclusion is correct, as we shall see shortly, it is premature: when the server's choice of direction of serve is not serially independent in point game i, 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 males and females and, indeed, different numbers of first serves for different players, a direct comparison of the *c.d.f.*s is not meaningful.

Comparing Male and Female Players

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.[21] Figure 11 shows the empirical densities of correlation coefficients for male and female tennis players for first serves. When computing the correlation coefficients for O'Neill's subjects we distinguish only between Joker and non-Joker

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

choices.



(a) Men (Hawk-Eye) (b) Women (Hawk-Eye) (c) O'Neill

Figure 11: 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.[22] In fact, the empirical *c.d.f.* of correlation coefficients for men first-order stochastically dominates the same *c.d.f.* for women.[23] For O'Neill's (1987) subjects, comprised of an unknown mixture of men and women, the mean is -0.113. Comparing students to tennis players, the difference of the means is not statistically for either men or women, likely a consequence of the small sample size for students.

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

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

[23] The null hypothesis that the correlation coefficients of males and females are drawn from the same distribution is decisively rejected by the two-sample KS test ($D = 0.1322$, KS p-value of 5.61×10^{-40}).

variations to cancel out for variation in the equilibrium mixture across point games.[24]

$Right_{t-1}$	-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.

6 Expertise and Conformity to Equilibrium

Receiver Expertise and Equality of Winning Probabilities

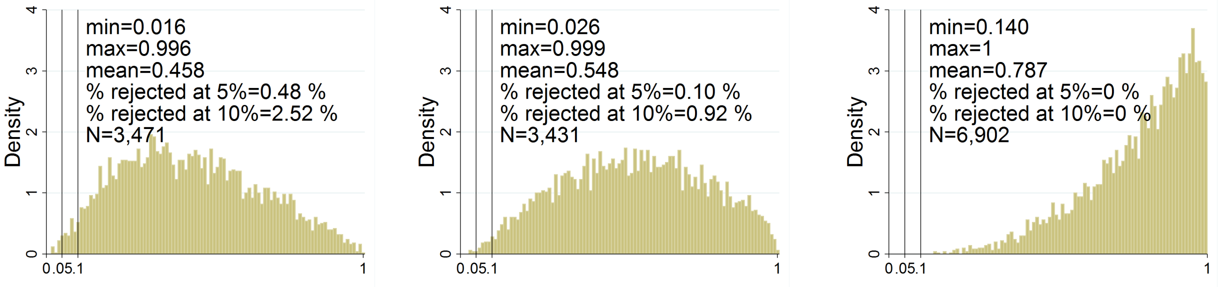
In Section 4 we established that the win rates of male professional tennis players conform remarkably closely to theory, while the win rates for women conform somewhat less closely. In other words, in men's tennis the receiver acts to equalize the server's winning probabilities, while this tends to be less true in women's tennis. In this section we consider whether the behavior of better (i.e., higher ranked) receivers 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 in this section is based on the subsample of matches for which we were able to obtain the receiver's ranking 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

[24] 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.

2006, only 69% of the point games for women.[25] The median rank for male players is 22 and for female players is 17.

According to the minimax hypothesis, the player receiving the serve acts to equalize the server's winning probabilities. Thus to evaluate the effect of expertise on behavior, we partition the data for first serves by men into two subsamples based on whether the *receiver* was a "top" player (i.e., ranked 17 or higher) or a "non-top" player (i.e., ranked below 17). The three panels of Figure 12 shows the empirical *c.d.f.*s of p-values when testing the joint null hypothesis of equality of winning probabilities for each subsample and for the sample of all point games for which we could obtain the receiver's rank.



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

Figure 12: KS test of $H_0 : p_L^i = p_R^i \ \forall i$ according to the receiver's rank (5000 trials)

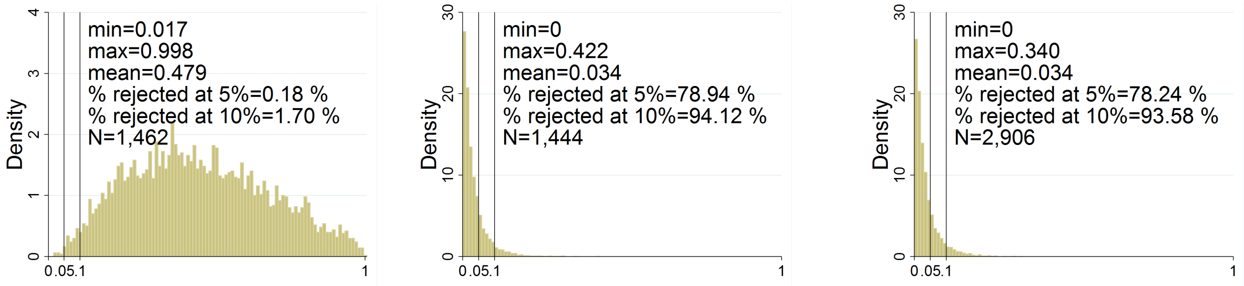
Panels (a) and (b) of Figure 12 show that the null hypothesis that winning probabilities are equalized is not rejected when servers face either top or non-top male receivers. The mean p-values are .458 and .548, respectively, and in only .48% and .10% of the trials is the null rejected at the 5% significance level. Hence we do not come close to rejecting the hypothesis that both top and non-top male receivers act to equalize the server's winning probability. This result matches our finding, reported in Figure 4, that winning probabilities for male players are equalized on first serves, consistent with the minimax hypothesis.[26]

Section 4 established (see Figures 6 and 7) that the behavior of female professional tennis players conformed less neatly to the minimax hypothesis. For first serves, the

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

[26] Figure 12(c) shows the empirical *c.d.f.* of p-values obtained when testing equality of winning probabilities on the sample of all 6,902 point games for which we could obtain the receiver's rank. It matches the empirical *c.d.f.* reported in Figure 4(b) for all 7,198 point games.

joint null hypothesis of equality of winning probabilities is rejected at the 5% level in 44.92% of all trials, and the mean p-value is .076. However, for the subsample of matches in which the receiver is ranked "top" (i.e., median or higher rank) the same null hypothesis does not come close to being rejected, as shown in panel (a) of Figure 13. In contrast, the null is decisively rejected for the subsample in which the receiver is ranked "non-top," as shown in panel (b). Panel (c) shows the null hypothesis is decisively rejected for the sample of all female players for whom we could obtain their rankings, and is the analogue of Figure 6(b) which reports the results for all female players.



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

Figure 13: KS test of $H_0 : p^i_L = p^i_R \; \forall i$ according to the receiver's rank

These results suggest that the best female players, when receiving the serve, do act to equalize the server's winning probabilities, in accordance with the minimax hypothesis.

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 smaller for women than for men. 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 will not be sufficiently successful to appear in our dataset. The serve in women's tennis 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. The best female players, nonetheless, do equalize the server's winning probabilities.

Server Expertise and Serial Correlation

In Section 5 we established that, inconsistent with the minimax hypothesis, both men and women exhibit negative serial correlation in the direction of serve. Furthermore, women exhibit more negative serial correlation than men.

Optimal play for the server requires that the direction of serve be serially independent, and this is true regardless of the receiver's rank. Here we show that higher ranked male players exhibit less serial correlation than lower ranked male players, while the degree of serial correlation does not depend on rank for women. This provides evidence that when there is a strong selection effect, as there is for men, then the behavior of better players conforms more closely to the minimax hypothesis.

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 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.

7 Discussion

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, as it is for our Hawk-Eye data, then the same test rejects the null even when it is true, and is thus not valid in our context. We show, in contrast, that when the empirical *c.d.f.* is generated from the randomized Fisher exact test t-values, then the test is valid, regardless of the number of point games.

We show further that the KS test based on the randomized Fisher t-values is more powerful than the tests used in the prior literature (the KS test based on the Pearson goodness of fit p-values and the Pearson joint test).[27] In Section 4 it was established that this more-powerful test, when applied to the WW data, confirms their original findings. (See Figure 3 and the associated discussion.) Here we show that the more powerful test, when applied to HHT's data, does not support their finding that the serve and return play of female professional tennis players and of players in junior matches is consistent with theory.

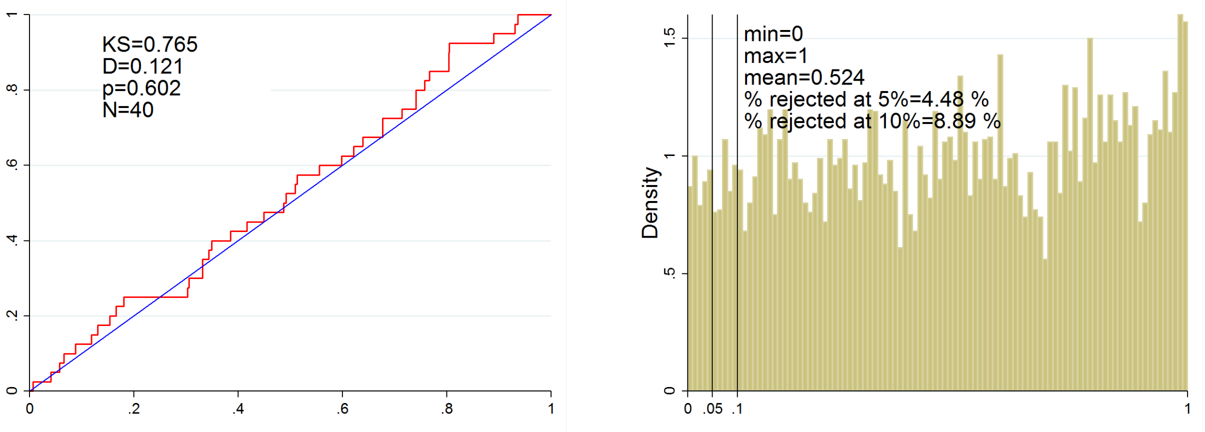
Valid Tests for Small and Large Samples

We first show that the KS test based on p-values from the Pearson goodness of fit test is valid for sample sizes of the kind studied in WW. A valid test generates p-values that are uniformly distributed when the null is true, and thus at the 5% significance level, say, rejects the null with probability .05.

The WW dataset had 40 point games, with an average of 75.65 serves per game, 54% of the serves were to the left, and the server's empirical winning probability was .64. We simulate data to roughly match these aggregate characteristics. In the simulated data there are 40 point games, every point game has 70 serves, each serve is

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

equally likely to be to the left or to the right, and a serve in either direction wins with probability 2/3. In particular, the null hypothesis of equality of winning probabilities is true for the simulated data.



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

Figure 14: KS test of $H_0 : p_L^i = p_R^i \; \forall i$ (Monte Carlo, 40 point games)

For each point game i, we compute the p-value from the Pearson goodness of fit test of the hypothesis that $p_L^i = p_R^i$. Under the null, the p-value is (asymptotically) uniformly distributed. Figure 14(a) shows an empirical *c.d.f.* of 40 such p-values obtained from simulating the WW data once. Applying the KS test to the empirical *c.d.f.* of p-values to test the null hypothesis that $p_L^i = p_R^i$ for all i yields a test statistic of $K = .765$ and associated p-value is .602.[28]

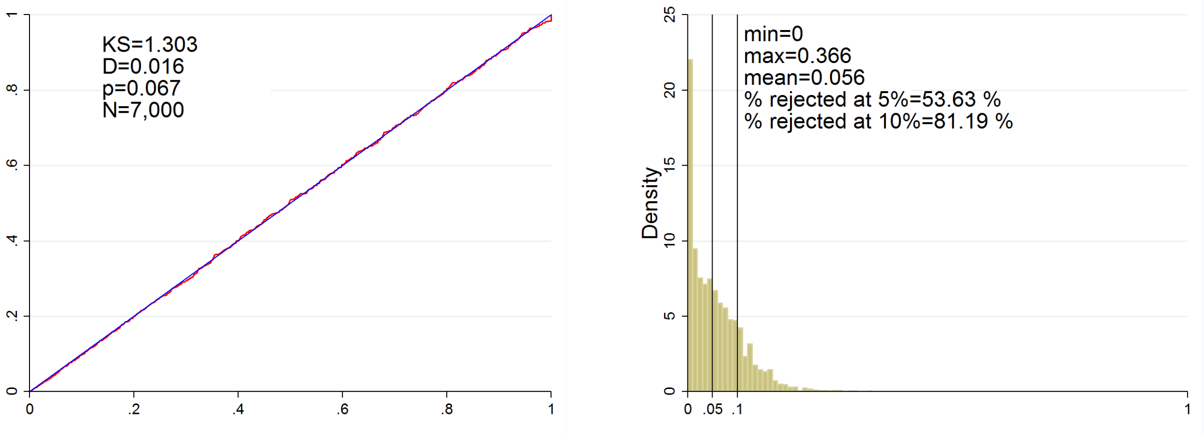
This KS test is valid if the associated p-values are uniformly distributed under the null. To verify that they are, we simulated the data 10,000 times, each time (i) generating an empirical *c.d.f.* of 40 p-values, (ii) applying the KS test to the empirical *c.d.f.* to determine whether the 40 p-values are uniformly distributed, and (iii) recording the associated p-value of the KS test.[29] Figure 14(b) shows that the empirical density of these p-values is indeed roughly uniform. At the 5% significance level, the (true) joint null hypothesis of equality of winning probabilities is rejected

[28] The KS test p-value is asymptotically uniformly distributed provided that each p^i in the empirical *c.d.f.* is an independent draw from the same continuous distribution. In this application, the Pearson p^i's are distributed $U[0,1]$ asymptotically (as the number of serves in point game i grows large).

[29] Importantly, a new simulated data set is created in each trial. For the densities provide in Section 4, by contrast, the data is fixed and the trials differ only in the realized t-values.

in 4.48% of the trials. Likewise, at the 10% significance level the joint null is rejected in 8.89% of the trials. Thus this test is valid for sample sizes of 40 point games.

Figure 15 reports the results of exactly repeating this process, but simulating data for 7000 point games (rather than 40) in order to match the size of the Hawk-Eye dataset. Figure 15(a) shows a representative empirical *c.d.f.* of 7000 p-values and Figure 15(b) shows the empirical density of the KS test p-values after 10,000 trials. It is immediately evident from Figure 15(b) that the empirical density of the KS test p-values is not close to uniform. At the 5% significance level, for example, we see that the KS test rejects the joint null hypothesis of equality of winning probabilities in the majority of trials (53.63%), even though it is true! This test is not, therefore, appropriate for datasets with a large number of point games.

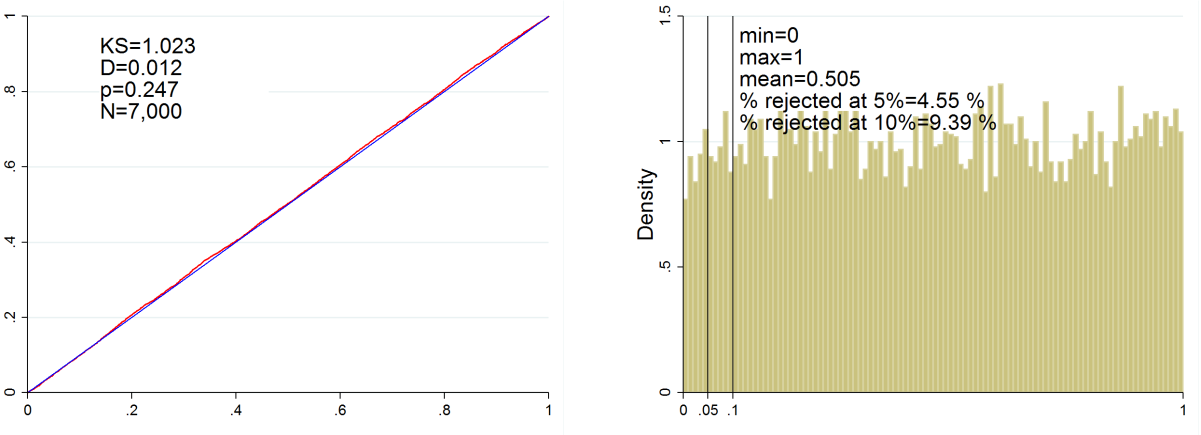


(a) An empirical *c.d.f.* of Pearson p-values (b) Density of KS test p-values (10,000 trials)

Figure 15: KS test of $H_0 : p^i_L = p^i_R \; \forall i$ (Monte Carlo, 7000 point games)

Why is the test valid when there are 40 points games, but not when there are 7000? Under the null hypothesis that $p^i_L = p^i_R$, the test statistic for the Pearson goodness of fit test is *asymptotically* distributed chi-square one as the number of serves in the point game grows large. Hence the p-values used to form the empirical *c.d.f.*s in Figures 14(a) and 15(a) are also only asymptotically uniformly distributed as well. When there are 7000 points games, the KS test is so powerful that it uncovers that the p-values are not truly distributed uniformly (since we observe only a finite number of serves in each point game). In particular, Figure 15(b) shows that when applying the KS test to empirical *c.d.f.*s comprised of 7000 of these p-values, we reject that they are uniformly distributed more often than not.

Figure 16 shows the results of simulating 7000 point games exactly as above, except that the KS test is based on the empirical distribution of the randomized Fisher exact test t-values. (As established in Section 4, the t-values are uniformly distributed under the null even for finite samples.) For each realization of 7000 t-values, we compute the associated p-value from the KS test. Figure 16(b) shows that the KS test performs as it should, with p-values that are distributed uniformly. The KS test rejects the (true) joint null hypothesis at the 5% significance level in 4.55% of 10,000 trials; at the 10% level it rejects the joint null hypothesis in 9.39% of the trials.



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

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

Figure 16: KS test of $H_0 : p^i_L = p^i_R \; \forall i$ (Monte Carlo, 7000 point games)

While we omit it here for brevity, Monte Carlo simulations show that the KS test based on the empirical distribution of t-values also performs as it should for small sample sizes as well.

For expositional convenience we have simulated data for *homogeneous* point games, i.e., every point game has 70 serves, serves left and right are equally likely, and the probability of winning a point is 2/3 for every serve. In Appendix B we show that the results reported in this subsection are robust to simulating data that matches, point game by point game, the observed characteristics of the WW data or the Hawk-Eye data.

The Power of Our Test

We have established that the KS test of the joint null hypothesis of equality of winning probabilities is valid for large samples when the empirical *c.d.f.* is based on the randomized Fisher exact test t-values, but not when it is based on the Pearson goodness of fit p-values. Using t-values rather than p-values also yields a test that is more powerful against the alternative hypothesis that the winning probabilities are unequal for left and right serves, as we now establish via Monte Carlo simulations.

To evaluate the power of our tests, we follow WW and frame our discussion in terms of the following hypothetical point game, where the entry in each cell is the probability that the server wins the point.

		Receiver		
		L	R	
Server	L	0.58	0.79	0.53 1/3
	R	0.73	0.49	0.46 1/3
		2/3	1/3	

In the game's mixed-strategy Nash equilibrium, the receiver chooses L with probability 2/3 and the server chooses L with probability 0.53 1/3. The probability the server wins the point is 0.65 for a serve in either direction. 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's equilibrium mixture equalizes 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

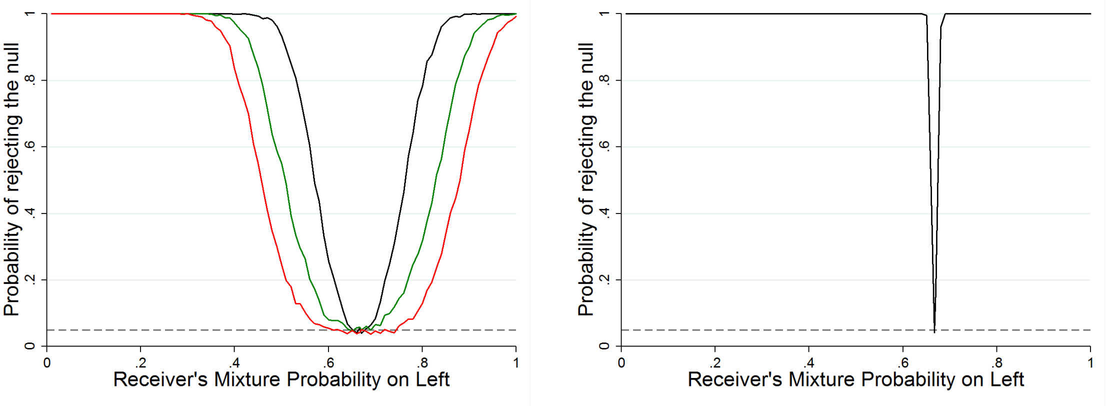
$$p_R(\theta) = .73\theta + .49(1 - \theta).$$

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.

We first simulate data for 40 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.[30] Figure 17(a) shows, as θ varies, the probability that the joint null

[30] To maintain conformity with the simulations discussed earlier in this section, we simulate the data under the assumption that serves in each direction are equally likely. Simulating it with the hypothetical point game's .53 1/3 equilibrium mixture probably on left has a negligible impact on the results.

hypothesis $H_0 : p^i_L = p^i_R \; \forall i \in \{1, \ldots, 40\}$ is rejected when $H_a(\theta) : p^i_L = p_L(\theta)$ and $p^i_R = p_R(\theta) \; \forall i \in \{1, \ldots, 40\}$ is true, for several different tests. The power function in red 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.[31] The power function in green 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 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 .256, while the Pearson joint test and the KS test based on the p's reject H_0 with probability .080 and .055, respectively.



(a) $N = 40$

(b) $N = 7000$

Figure 17: Power Functions for KS test based on t-values (black), p-values (red), and Pearson joint (green)

Figure 17(b) shows the power function for the KS test based on the randomized Fisher exact test t-values when data is simulated for 7000 point games, i.e., for approximately the number of point games in our Hawk-Eye data. 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 earlier, neither is a valid test (both tests reject the null with high probability even when it is true). Table 3 provides

[31] For each value of $\theta \in \{0, .01, .02, \ldots, .99, 1\}$ and for $\theta = 2/3$ the data is simulated 1000 times.

more detail about the power of the test for θ near its equilibrium value of 2/3.

True θ	KS based on t's	KS based on p's	Pearson joint test
0.65	0.995	0.554	0.275
0.66	0.454	0.548	0.241
2/3	0.042	0.515	0.197
0.67	0.169	0.564	0.229
0.68	0.960	0.551	0.244

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

The first row of Table 6 shows that that if $H_a(.65)$ 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 .995. Our more powerful test, coupled with a far larger dataset, yields a test of the joint null hypothesis of equality of winning probabilities that is far more powerful than any reported in the prior literature.

Re-Analysis of Prior Findings

WW found that the joint null hypothesis of equality of winning probabilities did not come close to being rejected. In Section 4 we established above that the same hypothesis is not rejected for the WW data even when using the more powerful KS test based on the randomized Fisher exact test t-values (see Figure 3 and the associated discussion).

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 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 for women and juniors is not robust to using the more powerful test based on the t-values.

Figure 18(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 5000 trials of the KS test based on the randomized Fisher t-values. The joint null hypothesis is not rejected once, even at the 10% level. Hence the more powerful test supports HHT's findings for men.

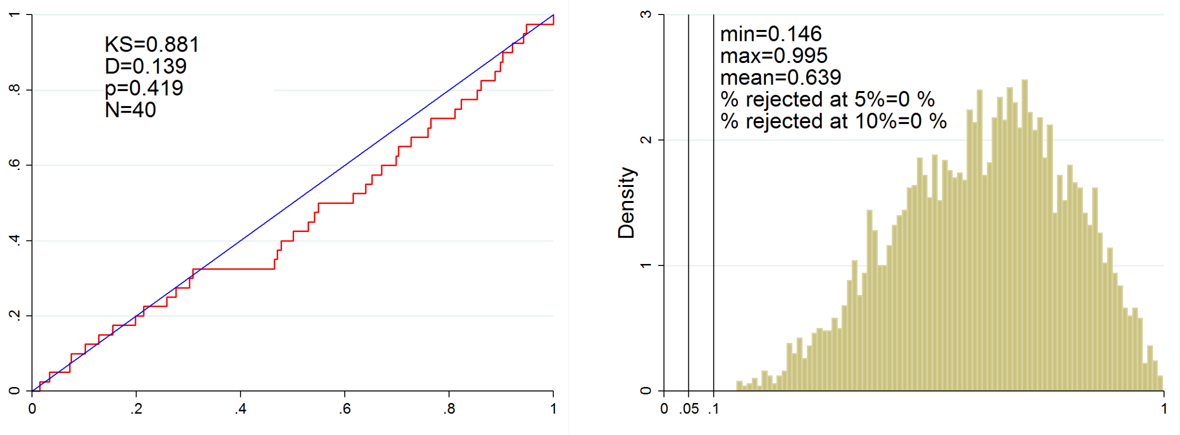


Figure 18(a): KS test for Men of $H_0 : p^i_L = p^i_R \; \forall i$ (HHT data)

Figures 18(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 18.48% of 5000 trials at the 5% level and in 45.30% 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 49.56% of the trials and at the 10% level in 77.48% of the trials.

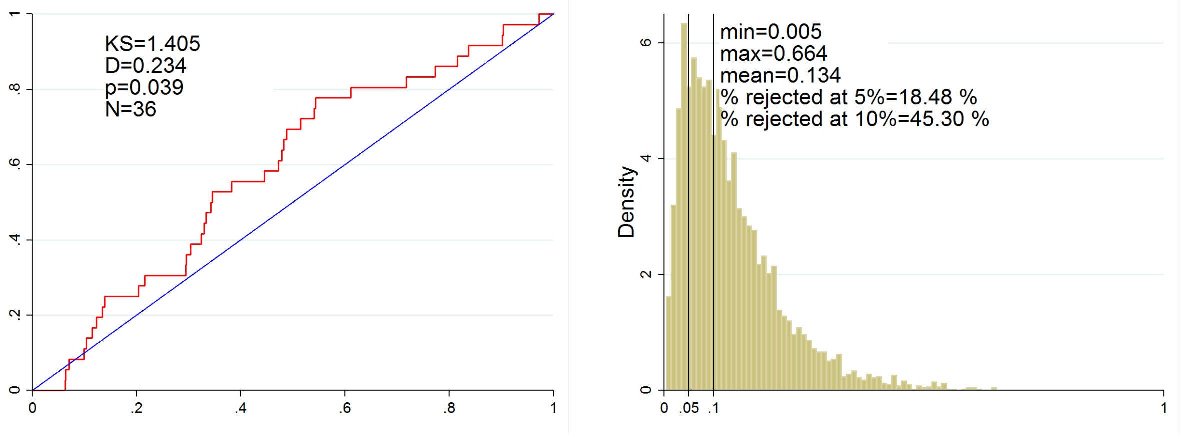


Figure 18(b): KS test for Women of $H_0 : p^i_L = p^i_R \; \forall i$ (HHT data)

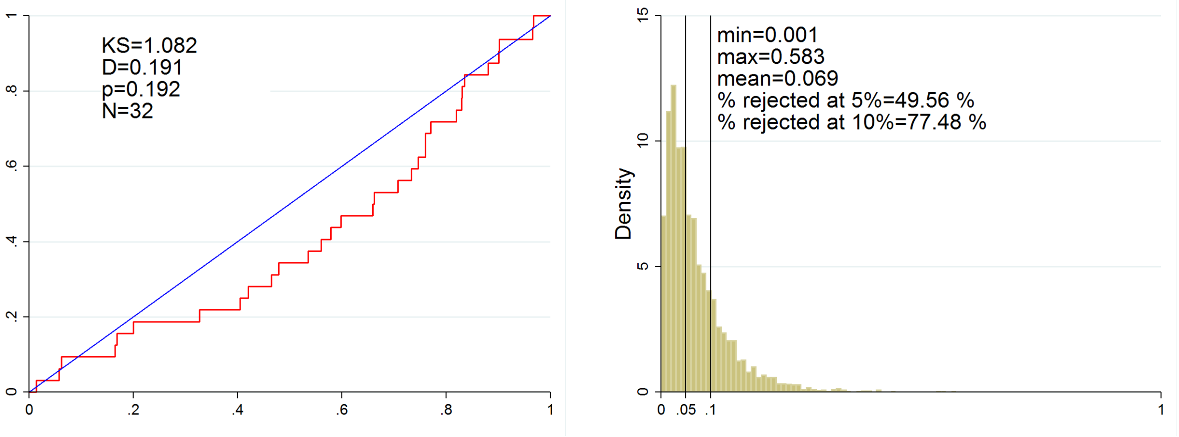


Figure 18(c): KS test for Juniors of $H_0 : p_L^i = p_R^i \; \forall i$ (HHT data)

Thus the greater power of the KS test based on the t-values changes the conclusions obtained from HHT's data.

8 Conclusion

Using data from professional tennis, the present paper provides by far the most powerful test of the minimax hypothesis hereto reported. The minimax hypothesis provides two testable predictions: (i) the probability that the server wins the point is the same for serves left and serves right, and (ii) the direction of serve is serially independent. The data provides remarkably strong support for the hypothesis that winning probabilities are equalized, especially for men. It also resoundingly rejects the hypothesis that the direction of the serve is serially independent. Behavior conforms more closely to the theory for men than women in both dimensions.

When the theory preforms poorly (e.g., equality of winning probabilities for women) or fails (e.g., serial independence for men), we provide evidence that the theory works better for more highly ranked players. To our knowledge the present paper is the first to provide field evidence that more skilled players behave in the field in closer conformity to the theory.

9 Appendix A

9.1 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, we drop those point games in which we observe 10 or fewer serves (Row (iv)).[32]

		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)	$\geq$ 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.

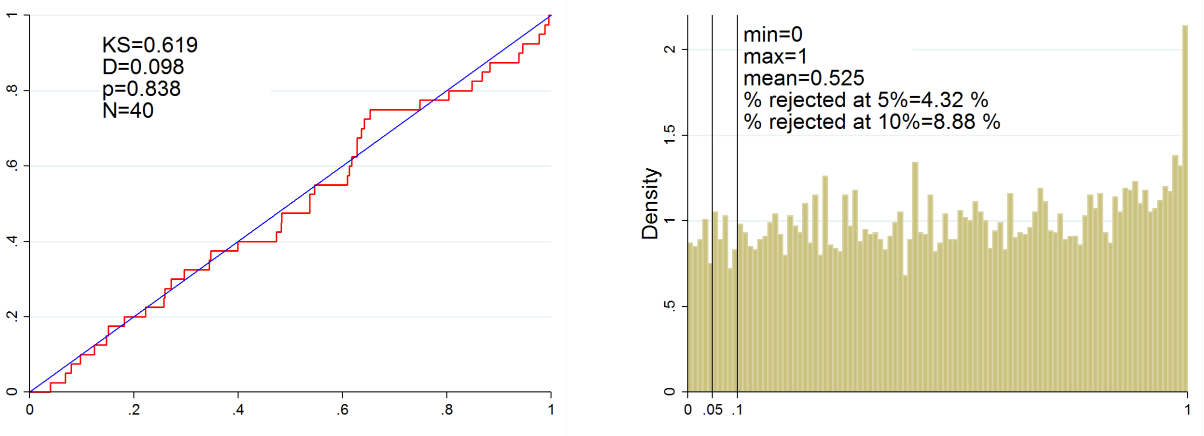
10 Appendix B: Not intended for publication

As a robustness check, here we reproduce the simulation results reported in Section 7, 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.

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

VALID TESTS FOR SMALL AND LARGE SAMPLES

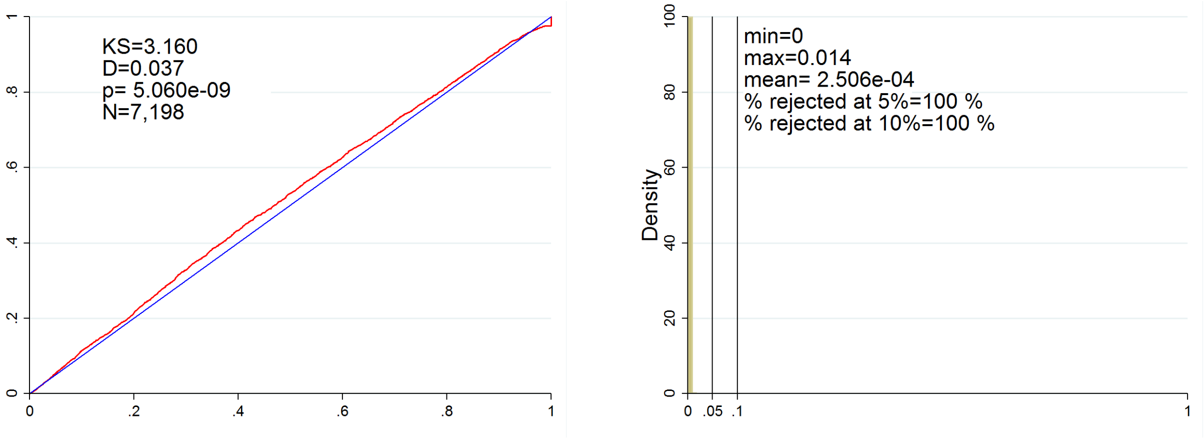
The subsection "Valid Tests for Small and Large Samples" in Section 6 established that the KS test of the null hypothesis of equality of winning probabilities ($p_L^i = p_R^i$ $\forall i$) based on the p-values from the Pearson goodness of fit test is valid for "small" samples (40 point games), but is not valid for "large" samples (7000 point games). Figure B1(b), the analogue to Figure 14(b), shows that the empirical distribution of p-values is approximately uniform when the simulated data matches the characteristics of WW data, point game by point game. Thus our conclusion in Section 6 that the KS test based on the p-values is a valid test for samples of the size studied in WW is robust to how the data is simulated,



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

Figure B1: KS test of $H_0 : p_L^i = p_R^i$ $\forall i$ (Monte Carlo, 40 point games)

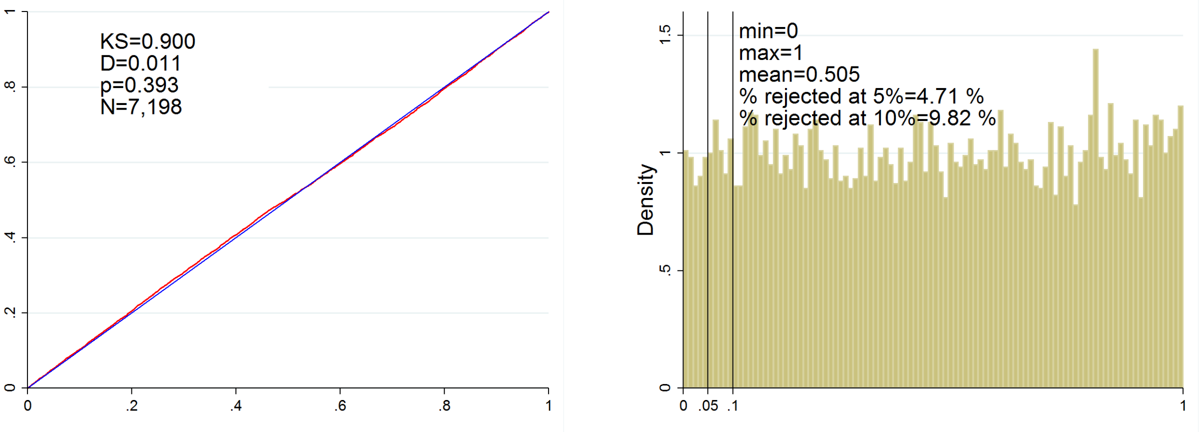
Figure B2(b) is the analogue of Figure 15(b). It shows that when data is simulated (under the null hypothesis) to match the Hawk-Eye data, then the KS test based on the p-values always rejects the null at the 5% significance level, and hence the test is invalid. In the simulation results report is Figure 15(b), the (true) null is rejected at the 5% level in only 53.63% of all trials. The failure of the KS test is even more striking in Figure B2 since the Hawk-Eye data has a smaller number of serves per point game – only 33 on average – than in the simulated data in Section 6, in which there were 70 serves per point game.



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

Figure B2: KS test of $H_0 : p_L^i = p_R^i \; \forall i$ (Monte Carlo, 7000 point games)

Figure B3(b) is the analogue of Figure 16(b). It shows that when data is simulated (under the null hypothesis) to match the Hawk-Eye data, then the KS test based on the t-values rejects the null at the 5% significance level in 4.71% of the trials. Moreover, it is visually evident that the empirical distribution of the KS test p-values is uniformly distributed, as it should be theoretically. Hence our conclusion in Section 6 that the KS test based on the t-values is a valid test is robust to simulating the data to match the Hawk-Eye data, point game by point game.



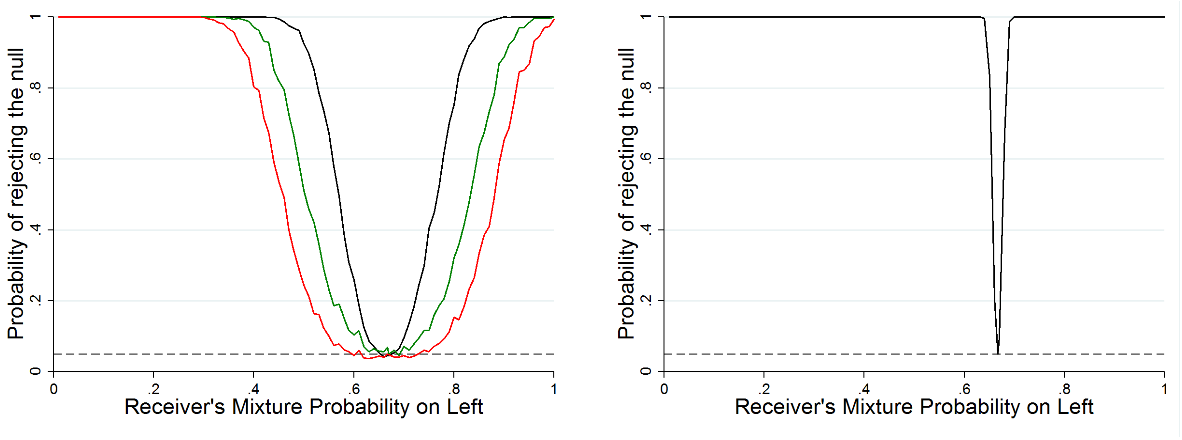
(a) An empirical *c.d.f.* of t-values (b) Density of KS test p-values (10,000 trials)

Figure B3: KS test of $H_0 : p_L^i = p_R^i \; \forall i$ (Monte Carlo, 7000 point games)

The Power of the 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 B4 is the analogue to Figure 17 and shows that the power functions in Figure 17 are largely unchanged when the data is simulated (under the null hypothesis) to match characteristic of the WW data (Figure B4(a)) or the Hawk-Eye data (Figure B4(b)).



(a) $N = 40$

(b) $N = 7198$

Figure B4: Power Functions for KS test based on t-values (black), p-values (red), and Pearson joint (green)

Table B1 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 B1 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 .713 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	0.833	1	0.755
0.66	0.205	1	0.729
2/3	0.050	1	0.713
0.67	0.098	1	0.744
0.68	0.671	1	0.732

Table B1: Rejection rate for H_0 at the 5% level, $N = 7198$

10.1 Ball Bounces

Figure B5 below shows actual and imputed ball bounces for male second serves from the deuce court.

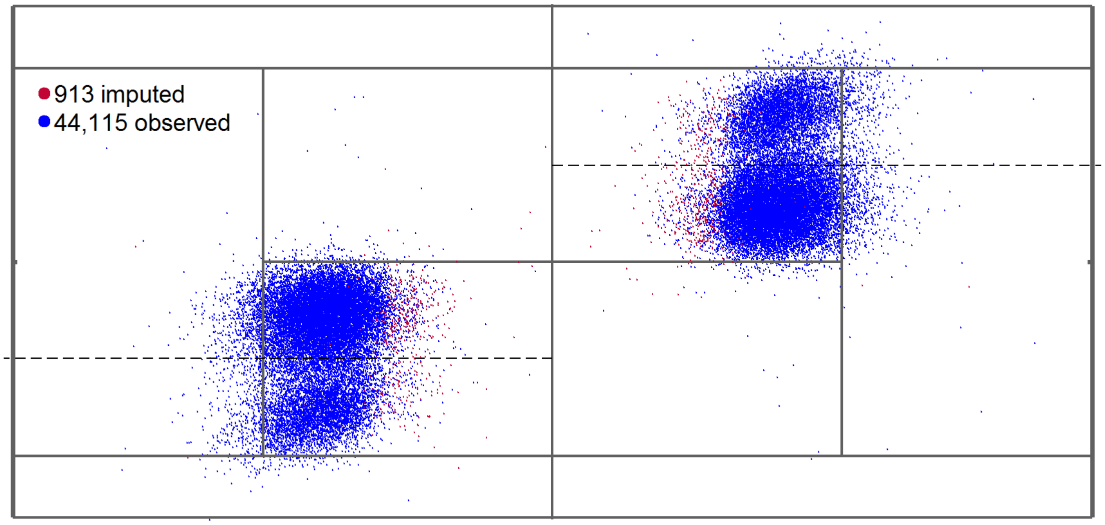


Figure B5: Ball Bounces for Deuce Court Second Serves by Men

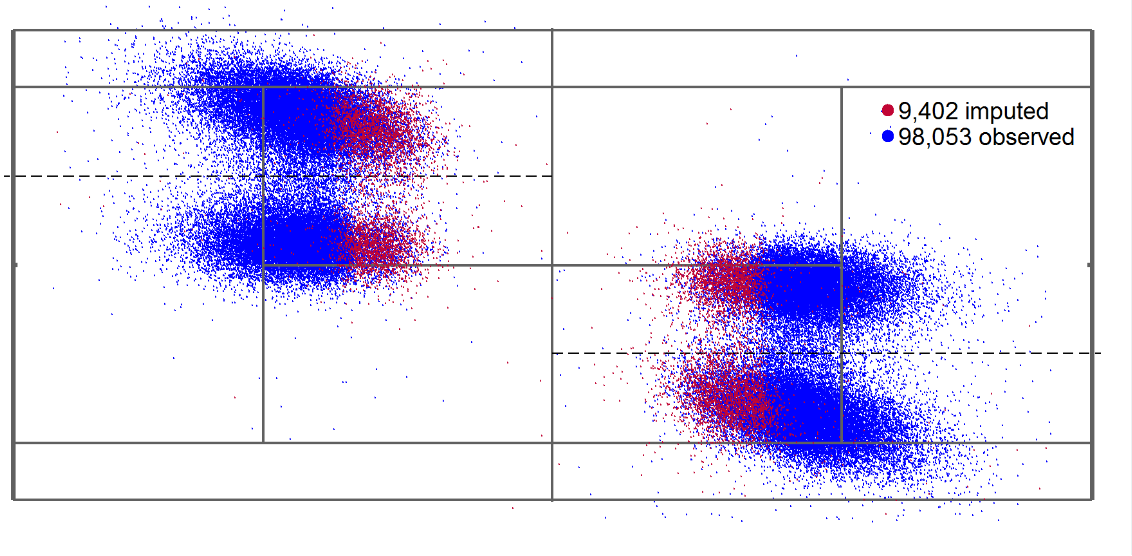


Figure B6: Ball Bounces for Ad Court First Serves by Men

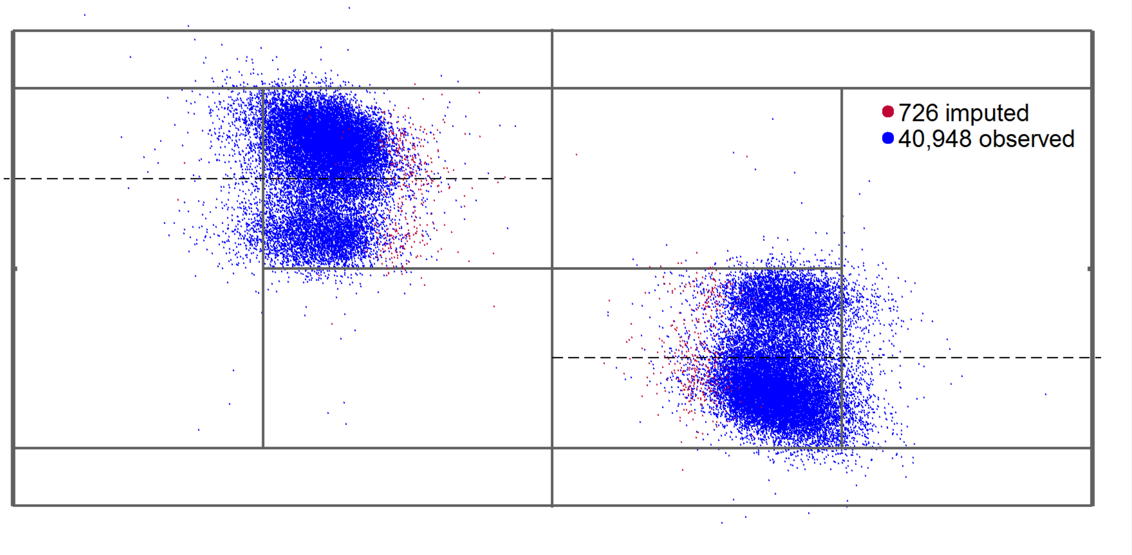


Figure B7: Ball Bounces for Ad Court Second Serves by Men

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

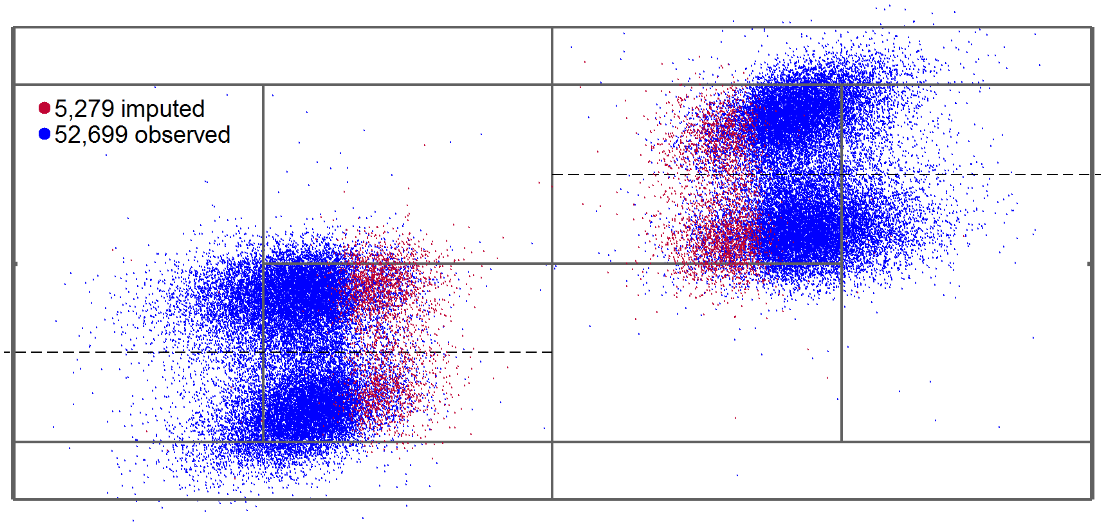


Figure B8: Ball Bounces for Deuce Court First Serves by Women

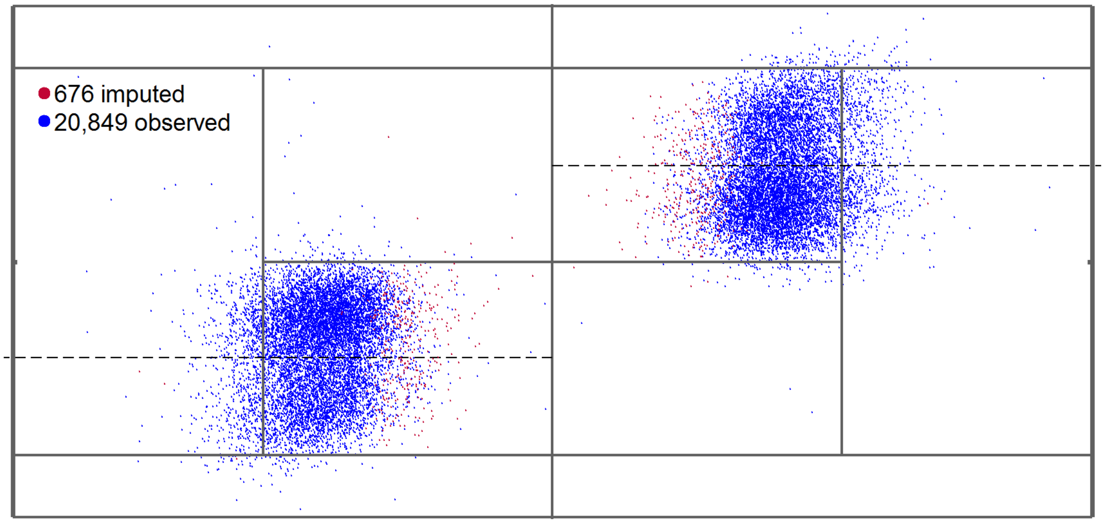


Figure B9: Ball Bounces for Deuce Court Second Serves by Women

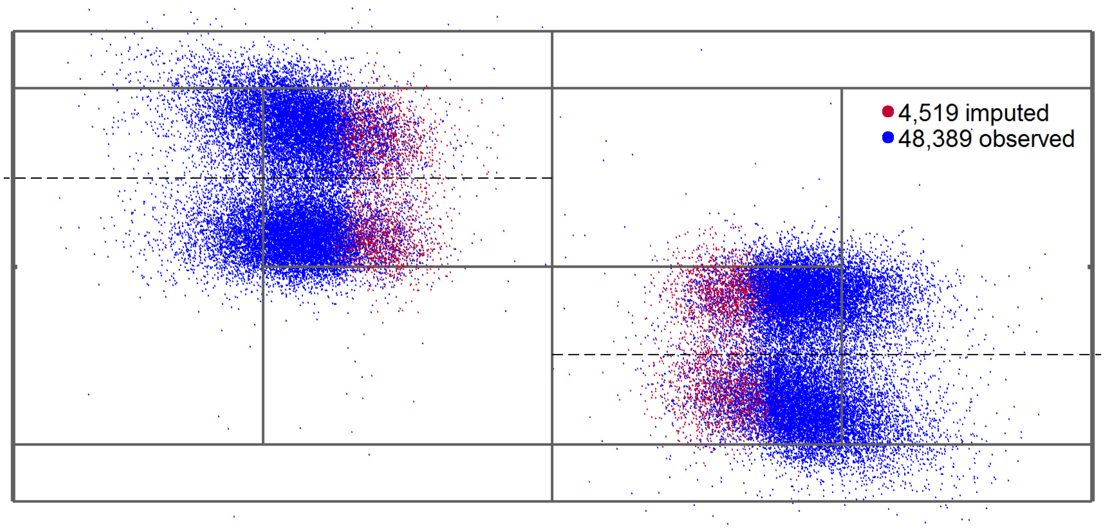


Figure B10: Ball Bounces for Ad Court First Serves by Women

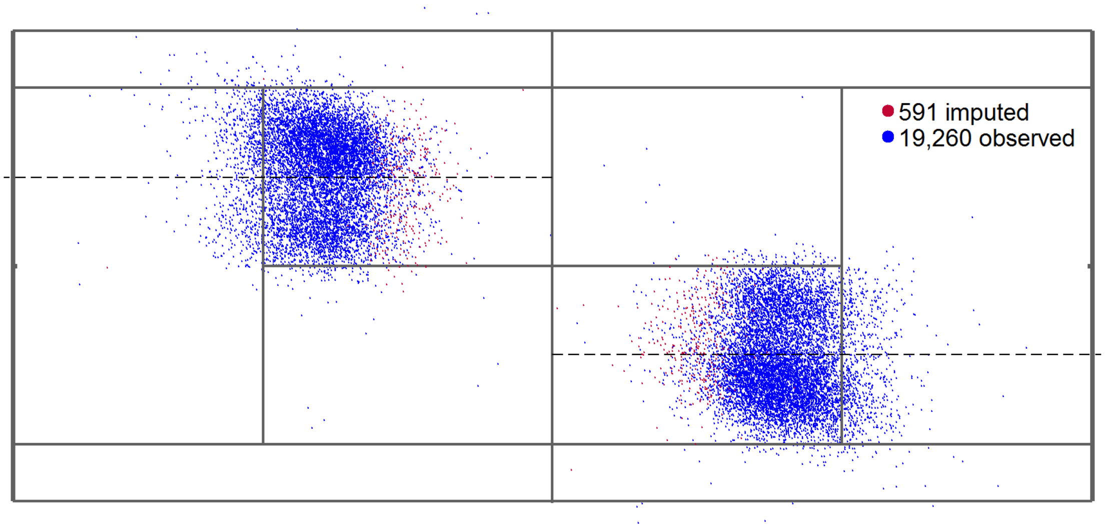


Figure B11: Ball Bounces for Ad Court Second Serves by Women

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] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

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

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

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

[14] 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.

[15] 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.

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

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

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

[19] 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.

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

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

[22] Wooders, J. (2008) "Does Experience Teach? Professionals and Minimax Play in the Lab," University of Arizona Working Paper #08-04.

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