Quantitative Economics 14 (2023), 981–1020 1759-7331/20230981

Expertise, gender, and equilibrium play

Romain Gauriot
Department of Economics, Deakin University

Lionel Page
School of Economics, University of Queensland

John Wooders
Division of Social Science, New York University Abu Dhabi and the Center for Behavioral Institutional Design

Mixed-strategy Nash equilibrium is the cornerstone of our understanding of strategic situations that require decision makers to be unpredictable. Using data from nearly half a million serves over 3000 tennis matches, and data on player rankings from the ATP and WTA, we examine whether the behavior of professional tennis players is consistent with equilibrium. We find that win rates conform remarkably closely to the theory for men, but conform somewhat less neatly for women. We show that the behavior in the field of more highly ranked (i.e., better) players conforms more closely to theory. We show that the statistical tests used in the prior related literature are not valid for large samples like ours; we develop a novel statistical test that is valid and show, via Monte Carlo simulations, that it is more powerful against the alternative that receivers follows a nonequilibrium mixture.

Keywords. Minimax, mixed strategy Nash equilibrium play, natural experiment.
JEL classification. C12, C15, C72.

1. Introduction

Laboratory experiments have been enormously successful in providing tightly controlled tests of game theory. The results of these experiments, however, have not been supportive of the theory for games with a mixed-strategy Nash equilibrium: student subjects do not mix in the equilibrium proportions and subjects exhibit serial correlation in their choices rather than the serial independence predicted by the theory. While the rules of an experimental game, which requires players to be unpredictable, may be simple to understand, it is far more difficult to understand how to play *well*. Student subjects

Romain Gauriot: romain.gauriot@deakin.edu.au
Lionel Page: lionel.page@uq.edu.au
John Wooders: john.wooders@nyu.edu
The authors are grateful to Guillaume Fréchette, Kei Hirano, Jason Shachat, and Mark Walker for useful comments. Wooders gratefully acknowledges financial support from the Australian Research Council's Discovery Projects funding scheme (Project DP140103566) and from Tamkeen under the NYU Abu Dhabi Research Institute Award CG005. Gauriot is grateful for financial support from the Australian Research Council's Discovery Projects funding scheme (Project DP150101307).

© 2023 The Authors. Licensed under the Creative Commons Attribution-NonCommercial License 4.0. Available at http://qeconomics.org. https://doi.org/10.3982/QE1563

no doubt understand the rules, but they have neither the experience, the time, nor the incentive to learn to play well. In professional sports, by contrast, players have typically devoted their lives to the game and they have substantial financial incentives, and thus it provides an ideal setting to test theory.

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

A critique of the results of prior studies using data from professional sports has been that they have low power to reject the theory.[1] Walker and Wooders (2001), henceforth WW, studies a data set comprised of approximately 3000 serves made in 10 men's championship tennis matches. Chiappori, Levitt, and Groseclose (2002), henceforth CLG, and Palacios-Huerta (2003), henceforth PH, study 459 and 1417 penalty kicks, respectively. Our data set, by contrast, contains the precise trajectory and bounce points of the tennis ball for nearly 500,000 serves from over 3000 professional tennis matches, and thereby provides an extremely powerful test of the theory. Camerer (2003) suggests that WW's focus on long matches, with the goal of generating a test with high statistical power, could introduce a selection bias in favor of equilibrium play. Our analysis does not suffer from this critique as it uses data from all the matches where the Hawk-Eye system was employed.

The large number of matches in our data set requires the development of a novel statistical test for our analysis. When the number of points played in each match is small relative to the overall number of matches, as it is in our data set, we show that a key statistical test employed in WW is not valid: even when the null hypothesis is true, the test rejects the null (implied by Nash equilibrium) that winning probabilities are equalized across directions of serve. By contrast, the test that we develop, based on the Fisher exact test, rejects the true null hypothesis with exactly probability α at the α significance level. We show via Monte Carlo simulations that our test, as an added bonus, is substantially more powerful than the test used in WW and the subsequent literature against the alternative that the receiver follows a nonequilibrium mixture.[2,3]

[1] See Kovash and Levitt (2009).

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

[3] It should be understood hereafter that all power comparisons are based on Monte Carlo simulations where the alternative hypothesis is that receivers follows a nonequilibrium mixture.

An unusual feature of our test is that the test statistic itself is random, and thus a different p-value is realized each time the test is conducted. It would be perfectly legitimate to run the test once and reject the null hypothesis if the p-value is less than the desired significance level. It is more informative, however, to report the empirical density of p-values obtained after running the test many times, and this is what we do. When reporting our results, we will make statements such as "the empirical density of p-values places an x% probability weight on p-values below 0.05." Reporting the empirical density reveals the sensitivity of our conclusions to the randomness inherent in the test statistic. Since randomized tests are seldom used, we complement our analysis with the implementation of a deterministic test in Appendix B.[4] The deterministic test has low power in comparison to our randomized test.

We find that the win rates of male professional tennis players are strikingly consistent with the equilibrium play. Despite the enormous power of our statistical test—due to the large sample size and the greater power of the test itself—we cannot 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 probability weight on p-values below 0.05 (i.e., the joint null hypothesis is never rejected at the 5% significance level). For second serves, it places almost no probability weight on p-values below 0.05.

The win rates for female players, by contrast, conform less neatly to theory. The empirical density function of p-values places a 44.73% weight on p-values below 0.05 for first serves, and a 16.1% weight on p-values below 0.05 for second serves. Nonetheless, the behavior of female professional tennis players over 150,000 tennis serves conforms far more closely to theory than the behavior of student subjects in comparable laboratory tests of mixed-strategy equilibrium. Applying our test to the data from O'Neill's (1987) classic experiment, for example, we obtain an empirical density function of p-values that places probability one on p-values less than 0.05. Hence, the null hypothesis that winning probabilities are equalized is resoundingly rejected based on the 5250 decisions of O'Neill's subjects while we obtain no such result for female professional tennis players, despite having vastly more data.

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

As just noted, win rates conform to equilibrium somewhat less neatly for women. Significantly, in women's matches in which the receiver is a "top" player, we do not come

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

close to rejecting equilibrium, while equilibrium is resoundingly rejected for the subsample in which the receiver is "nontop." This result shows that behavior of female receivers conforms more closely to theory for more highly ranked players.

What might explain this difference between male and female players? While the rules of the game are the same for men's and women's tennis, the payoffs in the contest for a point are different: 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.[5] Hence, it seems likely that the players' incentives to learn equilibrium, or selection effects that favor equilibrium play, differ for men and women. Given the greater speed of the serve, a receiver in men's tennis who fails to play equilibrium (and equalize the server's winning probabilities) may be more vulnerable to being exploited by the server. The serve is widely regarded as more important in men's than women's tennis (see Rothenberg (2017)).

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

Here, too, we find evidence that the behavior of higher ranked players conforms more closely to equilibrium: higher ranked male players exhibit less serial correlation in the direction of serve than lower ranked players.

Related literature

Our paper contributes to the literature investigating the degree to which the behavior of professions conforms to equilibrium. WW was the first paper to use data from professional sports to test the minimax hypothesis and the notion of mixed-strategy Nash equilibrium.[6] 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.

Hsu, Huang, and Tang (2007), henceforth HHT, broadens the analysis of WW. It found that win rates conformed to the theory for a sample of 9 women's matches, 8 junior's matches, and 10 men's matches. The greater power of our statistical test means that it potentially overturns these conclusions and indeed it does: Our test, applied to

[5]In other words, the value of the game for the point (for the server) is higher in men's tennis than women's tennis, which is likely driven by physical differences: men are taller and stronger, and thus deliver faster serves. In our data set, the average speed of the first serve is 160 kph for men and 135 kph for women. Only 0.45 seconds elapses between the serve and the first bounce in men's tennis.

[6]von Neuman's notion of Minimax, the foundation of modern game theory, and Nash equilibrium coincide in two-player constant-sum games, and we will use the terms minimax and equilibrium interchangeably.

HHT's data for women and juniors, puts weights of 18.1% and 49.2%, respectively, on p-values of less than 0.05. On the other hand, applying our test to WW's data or HHT's data for men, we reaffirm their findings that the behavior of male professional tennis players conforms to equilibrium. In both cases, the empirical density of p-values assigns zero probability to p-values below 0.05.

CLG studies a data set of every penalty kick occurring in French and Italian elite soccer leagues over a 3-year period (459 penalty kicks), and tests whether play conforms to the mixed-strategy Nash equilibrium of a parametric model of a penalty kick in which the kicker and goalkeeper simultaneously choose Left, Center, or Right. A challenge in using penalty kicks to test theory is that most kickers take few penalty kicks and, furthermore, a given kicker rarely encounters the same goalie. The latter is important since the contest between a kicker and goalie varies with the players involved, as do the equilibrium mixtures and payoffs.[7] CLG finds that the data conforms to the qualitative predictions of the model, for example, kickers choose "center" more frequently than goalies.[8] A key contribution of CLG is the precise identification of the predictions of equilibrium theory that are robust to aggregation across heterogeneous contests.

PH studies a group of 22 kickers and 20 goalkeepers who have participated in at least 30 penalty kicks over a 5-year period, in a data set 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 kickers.[9] 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 a period of several hours.[10] The relationship between the players' actions and the probability of winning the point is the same in every such instance, and thus the data from a single match can be used to test equilibrium theory. There is no need to aggregate data across matches and players as in CLG or PH.

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

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

[8] In a linear probability regression, the paper finds 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.

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

[10] 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.

play the centipede game much like typical student subjects. Our work differs by examining the effect of expertise on the conformity of behavior in the field to equilibrium play.

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

In Section 2, we present the model of a serve in tennis and the testable hypotheses implied by the theory. In Section 3, we describe our data. In Section 4, we describe our new statistical test of the hypothesis that winning probabilities are equalized, we present our results, and we show that the behavior of higher ranked players conforms more closely to theory than for lower ranked players. In Section 5, we report the results of our test that the direction of serve is serially independent. In Section 6, we compare the power of the WW test and the power of our new test, for the hypothesis that winning probabilities are equalized. We show that (i) the WW test is valid when the number of points in each match is large relative to the number of matches, but is not valid conversely, (ii) our new test is valid whether the number of matches is small (as in WW) or large, (iii) our new test is more powerful than the test used WW and subsequent studies, and we (iv) apply our test to the data from HHT.

2. Modeling the serve in tennis

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

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. We refer to a game of the kind in Figure 1 as a "point game."

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

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

Figure 1. The game for a point.

the serve alternates between serving from the ad court (the left side of the court) and from the deuce court (the right side). Since the players' abilities may differ when serving or receiving from one court or the other, the probability payoffs in Figure 1 may also depend upon whether the serve is from the ad or deuce court. At the first serve, the probability payoffs include the possibility that the server ultimately wins the point after an additional (second) serve. Since the second serve is the final serve, the probability payoffs for a second serve will be different than those for a first serve.[11]

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

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

A tennis match is a complicated extensive form game: The first player to win at least four points and to have won two more points than his rival wins a unit of scoring called a "game." The first player to win at least six games and to have won two games more than his rival wins a "set." In a five-set match, the first player to win three sets wins the match. The players, however, are interested in winning points only in so far as they are the means by which they win the match. The link between the point games and the overall match is provided in Walker, Wooders, and Amir (2011), which defines and analyzes a class of games (which includes tennis) called Binary Markov games. They show that Nash equilibrium (and minimax) play in the match consists of playing, at each point, the equilibrium of the point game in which the payoffs are the winning probabilities π_{sr} of Figure 1. Thus, play depends only on which player is serving, whether the point is an ad-court or a deuce-court point, and whether the serve is a first or second serve; it does not otherwise depend on the current score or any other aspect of the history of play prior to that point.[12]

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

[12]A binary Markov game consists of a binary scoring rule and a collection of point games. A binary scoring rule consists of (i) a finite set of states and (ii) two transition functions that govern how play proceeds from one state to the next. The states represent the possible scores of the match. There are two absorbing

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

FIGURE 2. An illustrative point game.

Testing the theory

Two testable implications come from the theory. The first is that a player obtains the same payoff from all actions which in equilibrium are played with positive probability. To illustrate this hypothesis, consider the illustrative point game in Figure 2.

The receiver's equilibrium mixture is to choose L with probability 2/3 and R with probability 1/3. When the receiver follows this mixture, then the server wins the point with the same probability (namely, 0.65) whether he serves L or he serves R.[13] In an actual tennis match, we do not observe the probability payoffs, and hence we cannot compute the receiver's equilibrium mixture. Nonetheless, so long as the receiver plays equilibrium, then the server's probability of winning the point will be the same for serves L and serves R.

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

The second implication of the theory, which comes from the equilibrium analysis of the extensive form game representing a match, is that the sequence of directions of the serve chosen by a server is serially independent. A server whose choices are serially correlated may be exploited by the receiver and therefore his play is suboptimal. To address this question, we focus on the rank of the server. Section 4 provides evidence that higher ranked male players exhibit less serial correlation in the direction of their serve than lower ranked males, and thus their behavior conforms more closely to theory.

3. THE DATA

Hawk-Eye is a computerized ball tracking system used in professional tennis and other sports to precisely record the trajectory of the ball. Our data set consists of the official

states, one where Player A has won the match and the other where Player B has won. From every state, there are two possible transitions: the state reached if Player A wins the current point and the state reached if Player B wins the point. Associated with each state is a normal form game—a "point game"—that governs how points are won at that state.

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

TABLE 1. Match characteristics.

		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

Hawk-Eye data for all matches played at the international professional level, where this technology was used, between March 2005 and March 2009.[14] Most of the matches are from Grand Slam and Association of Tennis Players (ATP) tournaments. Overall, the data set contains 3172 different singles matches. Table 1 provides a breakdown of the match characteristics of our data.

As the use of the Hawk-Eye system is usually limited to the main tournaments, the data set contains a large proportion of matches from top tournaments (e.g., Grand Slams). Within tournaments, the matches in our data set 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. Finally, the tournament structure of tennis means that top players appear in more matches in a tournament. As a consequence, the matches contained in the data set tend to feature the best male and female players.

For each point played, our data set records the trajectory of the ball, as well as the player serving, the current score, and the winner of the point.[15] When the server faults as a result of the ball failing to clear the net, then we extrapolate the path of the serve to identify where the ball would have bounced had the net not intervened. Figure 3 is a representation of a tennis court and shows the actual and imputed 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.

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

[15]Hawk-Eye 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.

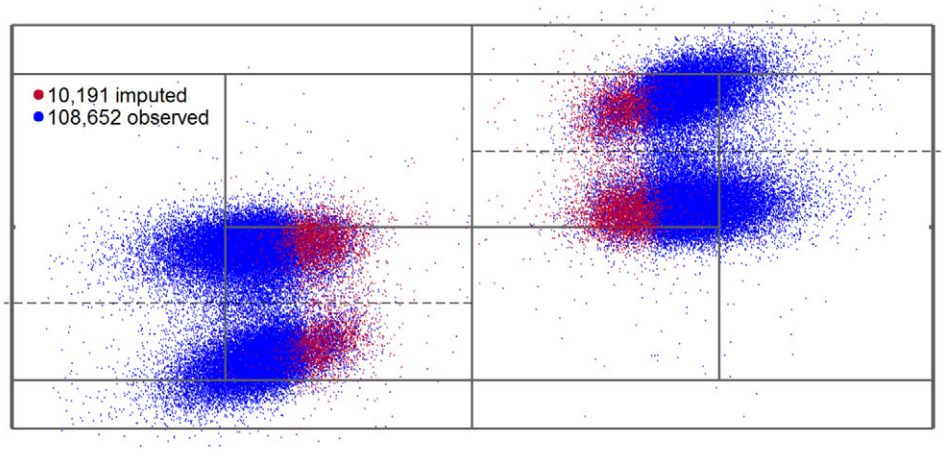


FIGURE 3. Ball bounces for deuce court first serves by men.

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

One could more finely distinguish serve directions, for example, 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 same theoretical winning probability.

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 E in the Online Supplementary Material (Gauriot, Page, and Wooders (2023)) for the ball bounces for second serves, and for serves from the ad court, and for serves by women.

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. We observe a total of 465,262 serves in the cleaned data. A detailed description of the data cleaning process is provided in Appendix A.

TABLE 2. Summary statistics.

Serve	Court	Serves	Point Games	Serves/Point Game	$P(L)$	$P(R)$	$P(Win\|L)$	$P(Win\|R)$
				Male				
First	Deuce	118,843	3615	32.88	45.07	54.93	65.02	64.71
First	Ad	107,455	3583	29.99	46.19	53.81	63.94	63.78
Second	Deuce	45,028	3615	12.46	30.02	69.98	52.53	52.02
Second	Ad	41,674	3583	11.63	33.45	66.55	51.99	52.72
				Female				
First	Deuce	57,978	2066	28.06	47.09	52.91	57.93	56.64
First	Ad	52,908	2042	25.91	52.08	47.92	56.55	55.97
Second	Deuce	21,525	2066	10.42	38.84	61.16	46.29	45.61
Second	Ad	19,851	2042	9.72	41.09	58.91	44.28	46.06

Table 2 reports summary statistics for the cleaned data, where $P(L)$ is the frequency of serves to the Left and $P(Win|L)$ is the frequency that the point is won following a serve to the left. Both male and female servers win the point more frequently on first serves than second serves. Men win both first and second serves more frequently than women.

4. TESTING FOR EQUALITY OF WINNING PROBABILITIES

Let p_j^i denote the true, but unknown, probability that the server in point game i wins the point when the first serve is in direction j. According to equilibrium theory, $p_L^i = p_R^i$ for each point game i, that is, the probability that the server wins the point is the same for serves left and for serves right in each point game.

Individual play and the Fisher exact test

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

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

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

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

In its standard application, the Fisher exact test rejects the null hypothesis at significance level α for n^i_{LS} such that $F(n^i_{LS}|n^i_S, n^i_L, n^i_R) \leq \alpha/2$ or $1 - F(n^i_{LS} - 1|n^i_S, n^i_L, n^i_R) \leq \alpha/2$.

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

We implement a randomized Fisher exact test of exactly size α as follows: For each point game i, let t^i be the random test statistic given by a draw from the uniform distribution $U[0, F(n^i_{LS}|n^i_S, n^i_L, n^i_R)]$ if n^i_{LS} takes its minimum value, that is, $n^i_{LS} = \min\{n^i_S - n^i_R, 0\}$, and by a draw from the distribution $U(F(n^i_{LS} - 1|n^i_S, n^i_L, n^i_R), F(n^i_{LS}|n^i_S, n^i_L, n^i_R)]$ otherwise. Under the null hypothesis that $p^i_L = p^i_R$, the test statistic t^i is distributed $U[0, 1]$.[16] Hence, rejecting the null hypothesis if $t^i \leq \alpha/2$ or $t^i \geq 1 - \alpha/2$ yields a test of exactly size α.

The Fisher exact test and the randomized Fisher exact test make the same (deterministic) decision for all realizations of n^i_{LS} except $n^i_{LS} = \bar{n}^i_{LS} + 1$ or $n^i_{LS} = \underline{n}^i_{LS} - 1$.[17] If $n^i_{LS} = \bar{n}^i_{LS} + 1$, then the Fisher exact test does not reject the null, while the randomized test rejects it with probability

$$\frac{\alpha/2 - F(\bar{n}^i_{LS}|n^i_S, n^i_L, n^i_R)}{F(\bar{n}^i_{LS} + 1|n^i_S, n^i_L, n^i_R) - F(\bar{n}^i_{LS}|n^i_S, n^i_L, n^i_R)}.$$

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

$$\frac{F(\underline{n}^i_{LS} - 1|n^i_S, n^i_L, n^i_R) - (1 - \alpha/2)}{F(\underline{n}^i_{LS} - 1|n^i_S, n^i_L, n^i_R) - F(\underline{n}^i_{LS} - 2|n^i_S, n^i_L, n^i_R)}.$$

It is the addition of randomization for these two realizations that yields a test of exactly size α.

Randomizing over whether to reject a specific null hypothesis might seem unnatural and, indeed, randomized tests are seldom used.[18] In our context, however, we are not interested in whether $p^i_L = p^i_R$ for any particular point game i, but rather whether the null hypothesis is rejected at the expected rate for the thousands of point games in our data set. Our use of the randomized Fisher exact test allows us to exactly control the size of the test and, therefore, the expected rejection rate under the null, without any appeal to an asymptotic distribution of a test statistic.

Table 3 shows the percentage of points games for which equality of winning probabilities is rejected for the Hawk-Eye data, for men and women and for both first and

[16] See Appendix C for the proof.

[17] Suppose $n^i_{LS} \leq \bar{n}^i_{LS}$. The randomized Fisher exact test rejects the null since $t^i \sim U(F(n^i_{LS} - 1|n^i_S, n^i_L, n^i_R), F(n^i_{LS}|n^i_S, n^i_L, n^i_R)]$ and, therefore, $t^i \leq F(n^i_{LS}|n^i_S, n^i_L, n^i_R) \leq F(\bar{n}^i_{LS}|n^i_S, n^i_L, n^i_R) \leq \alpha/2$.

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

TABLE 3. Rejection rate (Fisher exact test) for $H_0 : p_L^i = p_R^i$ (10,000 trials).

Setting	# Point Games	Significance Level	
		5%	10%
Men (1st Serve)	7198	5.06% (0.16)	10.01% (0.20)
Men (2nd Serve)	7198	5.02% (0.23)	10.13% (0.30)
Women (1st Serve)	4108	5.35% (0.22)	10.50% (0.28)
Women (2nd Serve)	4108	4.86% (0.30)	9.64% (0.40)

second serves. As just noted, for point game i the null hypothesis is rejected at the 5% significance level if either $t^i \leq 0.025$ or $t^i \geq 0.975$. Since t^i is random, each percentage is computed for 10,000 trials; the table reports the mean and standard deviation (in parentheses) of these trials. For men, for both first and second serves, the (mean) frequency at which the null is rejected at the 5% significance level is very close to 5%, the level expected if the null is true.[19] 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.

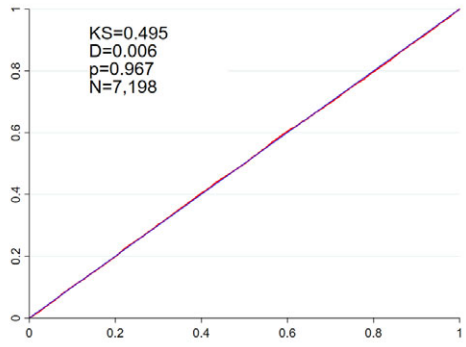
At the individual level, the rates at which equality of winning probabilities is rejected at the 5% and 10% significance level are consistent with the theory. These results are, however, only suggestive. In the next subsection, we report the results of our test of the hypothesis of interest, that $p_L^i = p_R^i$ for each point game i.

Aggregate play and the joint null hypothesis

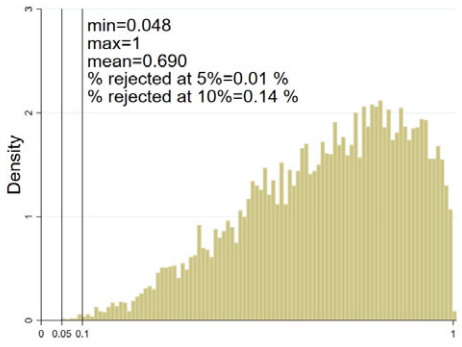
Of primary interest is the joint null hypothesis that $p_L^i = p_R^i$ for each point game i, and we use the t^i's generated from the randomized Fisher exact test to construct our test. Since the t^i's are independent draws from the same continuous distribution, namely the $U[0, 1]$ distribution, we can test the joint hypothesis by applying the Kolmogorov–Smirnov (KS) test to the empirical *c.d.f.* of the t-values. Formally, the KS test is as follows: The hypothesized *c.d.f.* for the t-values is the uniform distribution, $F(x) = x$ for $x \in [0, 1]$. The empirical distribution of N t-values, one for each point game, denoted $\hat{F}(x)$, is given by $\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} I_{[0,x]}(t^i)$, where $I_{[0,x]}(t^i) = 1$ if $t^i \leq x$ and $I_{[0,x]}(t^i) = 0$ otherwise. Under the null hypothesis, the test statistic $K = \sqrt{N}\sup_{x\in[0,1]} |\hat{F}(x) - x|$ has a known asymptotic distribution (see Mood, Graybill, and Boes (1974, p. 509)). Our appeal to an asymptotic distribution at this stage is well justified since there will be thousands of point games and associated t^i's for each of the joint null hypotheses we consider.

Figure 4(a) shows a realization of an empirical distribution (in red) of t-values for the Hawk-Eye data for first serves by men; the theoretical *c.d.f.* is in blue. The empirical and theoretical *c.d.f.*'s very nearly coincide. The value of the KS test statistic is $K = 0.495$

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



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



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

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

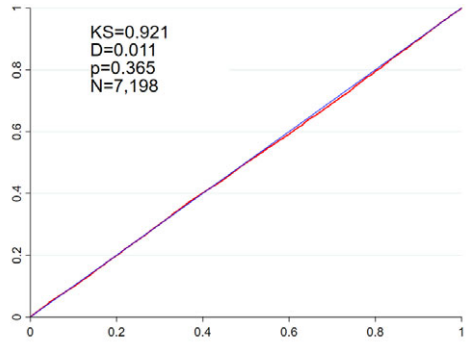
and the associated p-value is 0.967. The data is typical of the data that equilibrium play would produce: equilibrium play would generate a value of K at least this large with probability 0.967. Despite its enormous power, based on 226,298 first serves in 7198 point games, the test does not come close to rejecting the null hypothesis.

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

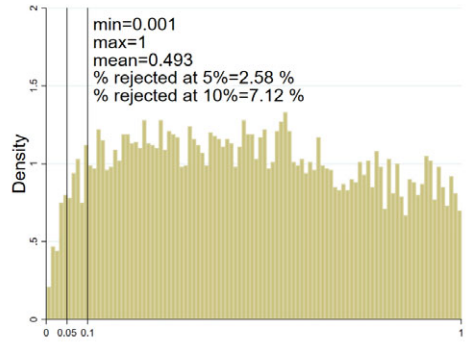
Figure 4(b) shows the empirical density of the KS test p-values obtained after 10,000 repetitions of the test. To construct the density, the horizontal axis is divided into 100 equal-sized bins $[0, 0.01]$, $[0.01, 0.02]$, ..., $[0.99, 1.0]$ and so, if 10,000 p-values were equally distributed across bins, then there would be 100 p-values per bin. The vertical height of each bar in the histogram is the number of p-values observed in the bin divided by 100. By construction, the area of the shaded region in Figure 4(b) is one, and hence it is an empirical density. The bins to the left of the vertical lines at 0.05 and at 0.10 contain, respectively, p-values for which the null is rejected at the 5% and 10% level.

Figure 4(b) shows that the conclusion above is indeed fully robust to the realizations of the t-values. The joint null hypothesis of equality of winning probabilities for first serves does not even come close to being rejected for the Hawk-Eye data for first serves by men. In only one instance, (0.01%) of 10,000 trials of the KS test is the null hypothesis rejected at the 5% level. In only 0.14% of the trials it is rejected at the 10% level. The mean p-value is 0.690, which is far from the rejection region.

Before proceeding, it is important to emphasize several aspects of our test. First, it is a valid test in the sense that if the null hypothesis is true (i.e., $p_L^i = p_R^i$ for each i), then the p-value obtained from the KS test is asymptotically uniformly distributed as the number of point games grows large. Second, conditional on any real-



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



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

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

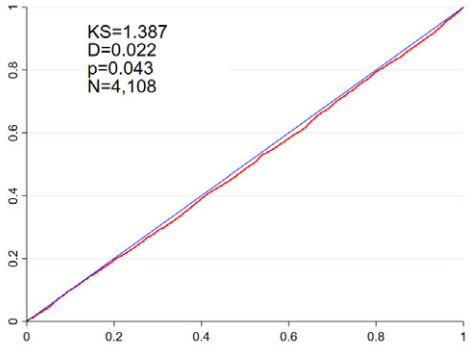
ization of the data, there is no reason to expect the KS test p-values obtained from running the test repeatedly (e.g., as in Figure 4(b)) to be uniformly distributed. Finally, as the number of serves in each point game grows large, then the intervals $U[F(n_{LS}^i - 1|n_S^i, n_L^i, n_R^i), F(n_{LS}^i|n_S^i, n_L^i, n_R^i)]$ from which the t-values are drawn shrink and the empirical density of the KS p-values collapses to a degenerate distribution.

Figure 5 shows the result of applying our test to the Hawk-Eye data for 86,702 second serves by men from 7198 point games. For a typical realization of the t-values, such as the one shown in Figure 5(a), the joint null hypothesis of equality of winning probabilities is not rejected. Figure 5(b) shows the density of KS test p-values after 10,000 trials. Only for a small fraction of these trials (2.58%) is the joint null rejected at the 5% level. For second serves as well, the KS test does not come close to rejecting the joint null hypothesis.[20]

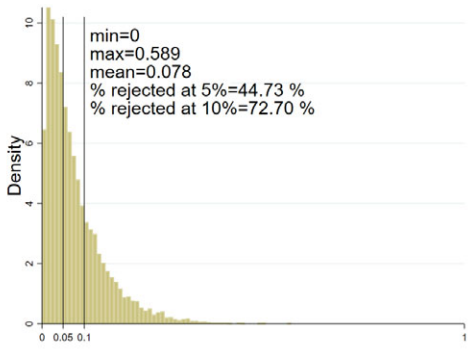
While the data for both first and second serves is strikingly consistent with the theory, comparing Figures 4(b) and 5(b) reveals that for second serves the conformity to theory is slightly less robust to the realization of the t^i's. This is a consequence of the fact that in tennis there are fewer second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

Our data also allows a powerful test of whether the play of women conforms to equilibrium. In the Hawk-Eye data for women, there are 110,886 first serves and 41,376 second serves, obtained in 4108 point games. For women, while the empirical and theoretical *c.d.f.*'s of t-values appear to the eye to be close, for many realizations of the t's the distance between them is, in fact, sufficiently large that the joint null hypothesis of equality of winning probabilities is rejected. Figures 6 and 7 show respectively the results of KS tests of the hypothesis that $p_L^i = p_R^i$ for all i, for first and second serves. For first

[20]While the numbers of point games for first and second serves are identical, there are fewer second serves than first serves in each point game and, therefore, the statistical power of the test is lower for second serves.



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



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

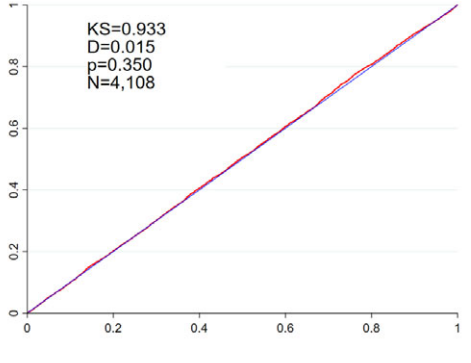
FIGURE 6. KS test for women of $H_0 : p_L^i = p_R^i \ \forall i$ (Hawk-Eye, First Serves).

serves, the null is rejected at the 5% and 10% significance level in 44.73% and 72.70% of 10,000 trials, respectively. In other words, run once, the test is nearly equally likely to reject the null as not at the 5% level; three out of four times the test rejects the null at the 10% level.

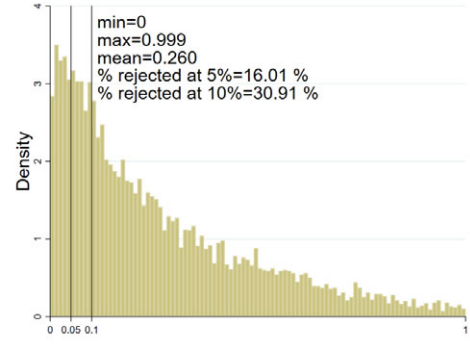
The results for second serves are more ambiguous. The null hypothesis tends not to be rejected at the 5% level: in only 16.01% of the trials is the p-value below 0.05.

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

The behavior of female professional tennis players, however, conforms far more closely to equilibrium than the behavior of student subjects in comparable laboratory tests of mixed-strategy Nash play. Figure 8(a) shows a representative empirical *c.d.f.* of 50 t-values obtained from applying our test to the data from O'Neill's (1987) classic experiment in which 50 subjects, in 25 fixed pairs, played a simple card game 105 times. In the game's unique mixed-strategy Nash equilibrium, the probability that player i wins

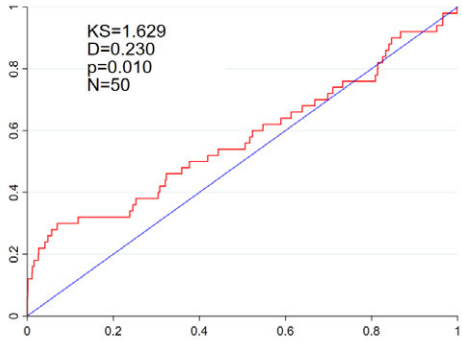


(a) An empirical *c.d.f.* of t-values

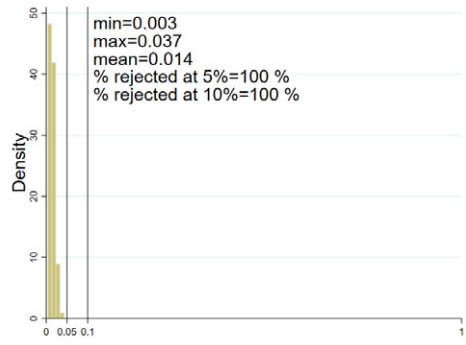


(b) Density of KS test p-values (10,000 trials)

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



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



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

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

a hand is the same when playing the joker card as when playing a number card (i.e., $p_J^i = p_N^i$). Nonetheless, for the empirical *c.d.f.* of t-values in Figure 8(a), the joint null hypothesis of equality of winning probabilities is decisively rejected, with a p-value of 0.01. 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.

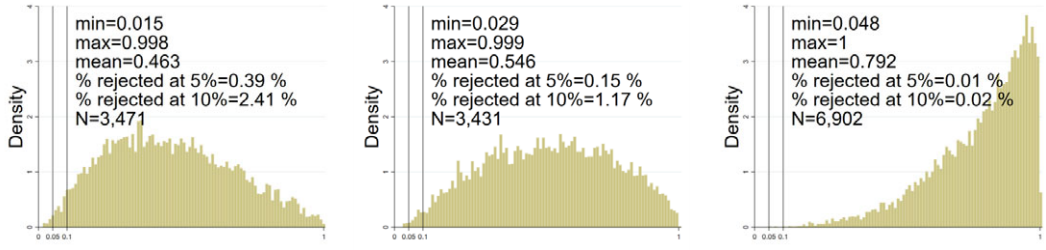
Hence, the behavior of female professional tennis players conforms far more closely to theory than the behavior of student subjects in laboratory experiments.

Expertise and equality of winning probabilities

Next, we consider whether the behavior of better (i.e., higher ranked) players conforms more closely to theory. The Association of Tennis Professionals (ATP) and the Women's Tennis Association (WTA) provide rankings for male and female players, respectively. Our analysis here is based on the subsample of matches for which we were able to obtain the receiver's rank at the time of the match. It consists of 96% of all point games for men, but since the ranking data was unavailable for women for the years 2005 and 2006, only 69% of the point games for women.[21] The median rank for male players is 22 and for female players is 17.

In a mixed-strategy equilibrium, the *receiver*'s play equalizes the server's winning probabilities. (In the illustrative example in Figure 2, the server's winning probability is 0.65 for each of his actions only if the receiver follows his equilibrium mixture.) Thus, to evaluate the effect of expertise on behavior, we partition the data for first serves into two subsamples based on whether the rank of the player receiving the serve was above or below the median rank. We say players with a median or higher rank are "top" players; all other players are "nontop." The three panels of Figure 9 show the empirical *c.d.f.*'s of KS-test p-values when testing the joint null hypothesis of equality of winning probabilities for each subsample of men and for the sample of all point games for which we could

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



(a) "Top" male receivers (b) "Nontop" male receivers (c) All male receivers

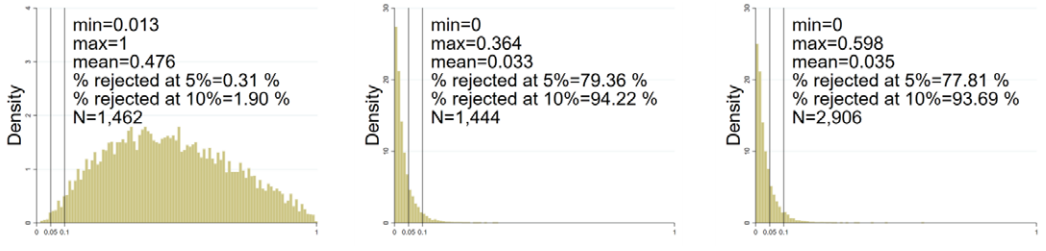
FIGURE 9. KS test for men of $H_0 : p_L^i = p_R^i \ \forall i$ by receiver's rank.

obtain the receiver's rank. Panel (c) is the analogue of Figure 4(b) and shows that the null of hypothesis that winning probabilities are equalized is not rejected for the sample of 6902 point games for which we have the receiver's rank.

Panels (a) and (b) of Figure 9 show that the null hypothesis that winning probabilities are equalized is not rejected when servers face either top or non-top male receivers. In only 0.39% and 0.15% of the trials is the null rejected at the 5% significance level; p-values are typically far from the rejection region. Hence, we do not come close to rejecting the hypothesis that male receivers act to equalize the server's winning probability, for either top or nontop receivers. This result is not surprising given the close conformity of the data to the theory on the whole sample.

Figures 6 and 7 established that the behavior of female professional tennis players conformed less neatly to equilibrium. For first serves, the joint null hypothesis of equality of winning probabilities is rejected at the 5% level in 44.73% of all trials. Figure 10(c) shows that a similar conclusion holds for the subsample of 2906 point games for which we were able to obtain the receiver's rank.

Figures 10(a) and (b) show a striking difference between the play of top and nontop female receivers: for the subsample of matches in which the receiver is ranked "top" the joint null hypothesis of equality of winning probabilities does not come close to being rejected, as shown in panel (a). In contrast, the null is decisively rejected for the subsample in which the receiver is ranked "nontop," as shown in panel (b). The best female players, when receiving the serve, do act to equalize the server's winning probabilities, in



(a) "Top" female receivers (b) "Nontop" female receivers (c) All female receivers

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

accordance with equilibrium. To our knowledge, this is the first evidence in the literature showing that behavior in the field of better players conforms more closely to theory.

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

Comparing tests

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

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

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

In Section 6, we verify via Monte Carlo simulations that the WW test is not valid when the number of point games is large relative to the number of serves. We show it is valid when the number of point games is not large relative to the number of serves, at it was for the WW data set of 40 point games.[22] Monte Carlo simulations show that our test is more powerful than the WW test when the WW test is valid. Our test, therefore, has two significant advantages over the WW test: (i) it is valid even when the number of point games is large, and (ii) it is more powerful than the WW test.

[22]In favor of the WW test, it makes a deterministic decision—it either rejects the null or not—and hence the results are easier to interpret, even if it falsely rejects a true null when the number of point games is large.

5. Serial independence

We test the hypothesis that the server's choice of direction of serve is serial independent. For each point game i, let $s^i = (s^i_1, \ldots, s^i_{n^i_L+n^i_R})$ be the sequence of first-serve directions, in the order in which they occurred, where $s^i_j \in \{L, R\}$ is the direction of the jth serve. Let r^i denote the number of runs in s^i. (A run is a maximal string of identical symbols, either all L's or all R's.) Under the null hypothesis of serial independence, the probability that there are exactly r runs in a randomly ordered list of n^i_L occurrences of L and n^i_R occurrences of R is

$$f_R(r|n^i_L, n^i_R) = \begin{cases} 2\binom{n^i_L-1}{r/2-1}\binom{n^i_R-1}{r/2-1} \Big/ \binom{n^i_L+n^i_R}{n^i_L} \\ \quad \text{if } r \text{ is even,} \\ \binom{n^i_L-1}{(r-1)/2}\binom{n^i_R-1}{(r-3)/2} + \binom{n^i_L-1}{(r-3)/2}\binom{n^i_R-1}{(r-1)/2} \Big/ \binom{n^i_L+n^i_R}{n^i_L} \\ \quad \text{if } r \text{ is odd,} \end{cases}$$

for $r \in \{2, \ldots, n^i_L + n^i_R\}$. Let $F_R(r|n^i_L, n^i_R)$ be the associated *c.d.f.* At the 5% significance level, the null is rejected if $F_R(r|n^i_L, n^i_R) \leq 0.025$ or if $1 - F_R(r-1|n^i_L, n^i_R) \leq 0.025$, that is, if the probability of r or fewer runs is less than 0.025 or the probability of r or more runs as less than 0.025. In the former case, the null is rejected since there are too few runs, that is, the server switches the direction of serve too infrequently to be consistent with randomness. In the latter case, the null is rejected as the server switches direction too frequently.

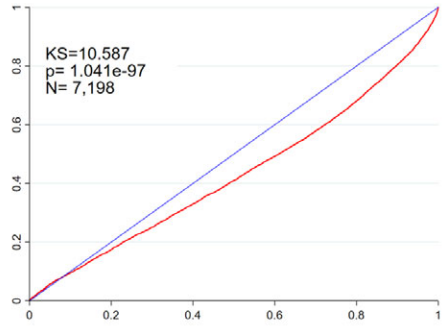
To test the joint null hypothesis that first serves are serially independent, for each point game i we draw the random test statistic t^i from the $U[F_R(r^i - 1|n^i_L, n^i_R), F_R(r^i|n^i_L, n^i_R)]$ distribution. Under the joint null hypothesis of serial independence, each t^i is distributed $U[0,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

Figure 11 shows representative empirical *c.d.f.*'s of t-values for first serves (left panel) and for second serves (right panel) for the Hawk-Eye data for men. The KS test rejects the joint null hypothesis of serial independence, for both first and second serves, with p-values virtually equal to zero.[23] 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 frequent switching, that is, there are more than the expected number of large t-values.

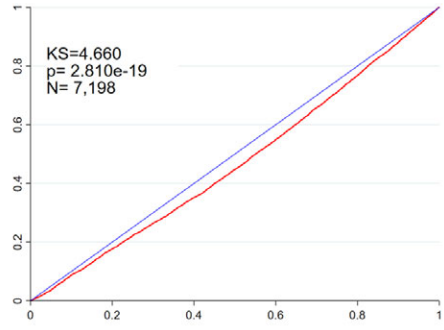
The empirical density of the KS test p-values that we have provided in prior figures is omitted for Figure 11 since the p-values are virtually zero for every realization of the t's.

Figure 12 shows representative empirical *c.d.f.*'s of t-values for first and second serves by women for the Hawk-Eye data. Women also exhibit negative serial correlation

[23]At the individual player level, serial independence is rejected in point game i at the 5% significance level if $t^i \leq 0.025$ or $t^i \geq 0.975$. For first serves, we reject serial independence as a result too few runs (i.e., $t^i \leq 0.025$) for 2.9% of the point games, and reject it as a result of too many runs (i.e., $t^i \geq 0.975$) for 7.0% of the point games.



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



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

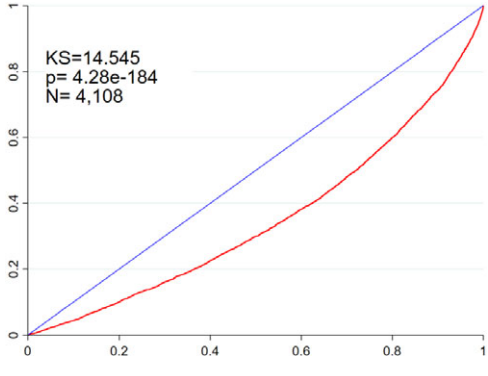
FIGURE 11. KS test for men of $H_0 : s^i$ is serial independent $\forall i$ (Hawk-Eye).

in the direction of serve, for both first and second serves, with the null of serial independence rejected at virtually any significance level.

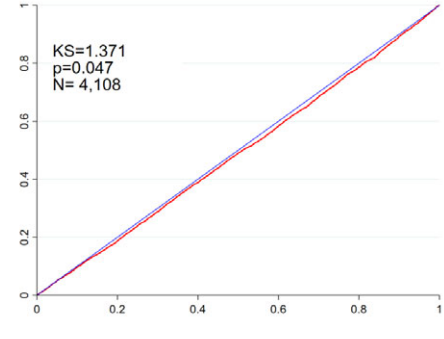
Comparing Figures 11(a) and 12(a), one might be tempted to conclude that women exhibit more serial correlation in first serves than men since the empirical *c.d.f.* of t-values is further from the theoretical one (namely, the 45-degree line) for women. While this conclusion is correct, as we shall see shortly, it is premature: when the server's choice of direction of serve is not serially independent in point game i, then the distribution of t^i will tend to depend on the number of first serves. Since we observe different numbers of first serves for men and women and, indeed, different numbers of first serves for different players, a direct comparison of the *c.d.f.*'s is not meaningful.

Gender and serial correlation

To determine the degree of serial correlation in first serves, and whether the difference between male and female players is statistically significant, we compute, for every point game, the Pearson product-moment correlation coefficient between successive

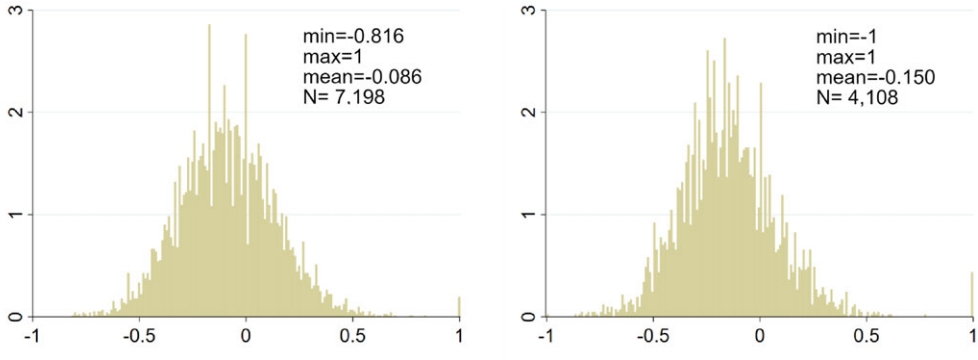


(a) First Serve: Empirical *c.d.f.* of t-values



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

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



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

FIGURE 13. Empirical density of correlation coefficients, first serves.

serves.[24] Figure 13 shows the empirical densities of correlation coefficients for male and female tennis players for 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.[25]

Table 4 shows the result of a logit regression for first serves in which the dependent variable is the direction of the current serve and the independent variables are the direction of the prior serve (from the same point game) and the direction of the prior serve interacted with gender. We use a fixed effect logit, using only within point game variation, to cancel out variation in the equilibrium mixture across point games.[26]

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

TABLE 4. Serial correlation and gender.

$Right_{t-1}$	-0.659
S.E.	(0.014)
$Right_{t-1} \times male$	0.329
S.E.	(0.017)
N_{serves}	325,394
Fixed effect	point game

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

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

[26]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.

with probability 0.418 if the server is male but will be right with probability only 0.341 if the server is female.[27]

Expertise and serial correlation

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

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.[28] 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

TABLE 5. Serial correlation and player rank.

	Men	Women
$Right_{t-1}$	−0.577	−0.689
S.E.	(0.027)	(0.053)
$Right_{t-1} \times \tilde{R}_{\text{server}}$	0.067	0.008
S.E.	(0.005)	(0.008)
$Right_{t-1} \times \tilde{R}_{\text{receiver}}$	−0.002	0.004
S.E.	(0.005)	(0.008)
N_{serve}	207,418	77,508
$N_{\text{pointgame}}$	6887	2901
Fixed effect	point game	point game

[27]In particular, on the next serve the probability of a serve to the right is, for males and females, respectively,

$$\Pr(Right_t | Right_{t-1}, male) = \frac{\exp(-0.659 + 0.329)}{1 + \exp(-0.659 + 0.329)} = 0.418,$$

and

$$\Pr(Right_t | Right_{t-1}, female) = \frac{\exp(-0.659)}{1 + \exp(-0.659)} = 0.341.$$

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

ranked players have *higher* values of $\tilde{R}$, for example, the players ranked first, second, third have values of $\tilde{R}$ equal to 8, 7, and 6.415, respectively.[29]

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

6. Statistical power and a reevaluation

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 point games is large, then the same test rejects the null even when it is true, and is thus not valid. We show, by contrast, that if the empirical *c.d.f.* is generated from the randomized Fisher exact test t-values as we propose, then the test is valid even when the number of point games is large.

We show further that our KS test based on the randomized Fisher t-values is more powerful than the two tests used in the prior literature: the KS test based on the Pearson goodness-of-fit p-values and the Pearson joint test.[30] A more powerful test has the potential to reverse the conclusions from WW (for men) and HHT (for men, women, and juniors) that the win rates in the serve and return play of professional tennis players are consistent with equilibrium. We show that the conclusions of WW and HHT for men are robust. However, the more powerful test does not support HHT's finding that the serve and return play of female professional tennis players and of players in junior matches is consistent with theory.

The power of our test

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

$$p_L(\theta) = 0.58\theta + 0.79(1-\theta)$$

[29] As described in Klaassen and Magnus (2014, pp. 107–110), the measure $\tilde{R}(\text{ATP rank}) = 8 - \log_2(\text{ATP rank})$ can be interpreted as a (smoothed) measure of the "expected" round a player reaches in a tennis tournament. A player ranked 1 has $\tilde{R} = 8$, that is, he reaches the final round (round 7) and wins. A player ranked 2 has $\tilde{R} = 7$, that is, he reaches the final round and loses. A player ranked 4 has $\tilde{R} = 6$, that is, he reaches the semifinal and loses, and so on. A player ranked 128 has $\tilde{R} = 1$, that is, he enters the tournament and does not advance to the next round.

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

TABLE 6. Rejection rate for H_0 at the 5% level, $N = 7000$.

True θ	KS based on t's	KS based on p's	Pearson joint test
0.65	0.850	1.00	0.692
0.66	0.221	1.00	0.650
2/3	0.047	1.00	0.654
0.67	0.089	1.00	0.643
0.68	0.662	1.00	0.682

and

$$p_R(\theta) = 0.73\theta + 0.49(1-\theta).$$

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

Since we have data for 7198 point games for men, we first simulate data for 7000 point games with payoffs as given above. In the simulated data every point game has 30 serves, and serves in each direction are equally likely.[31] Table 6 shows, as θ varies near its equilibrium value of 2/3, the probability that the joint null hypothesis $H_0 : p_L^i = p_R^i$ $\forall i \in \{1, \ldots, 7000\}$ is rejected at the 5% significance level when $H_a(\theta) : p_L^i = p_L(\theta)$ and $p_R^i = p_R(\theta)$ $\forall i \in \{1, \ldots, 7000\}$ is true, for several different tests.

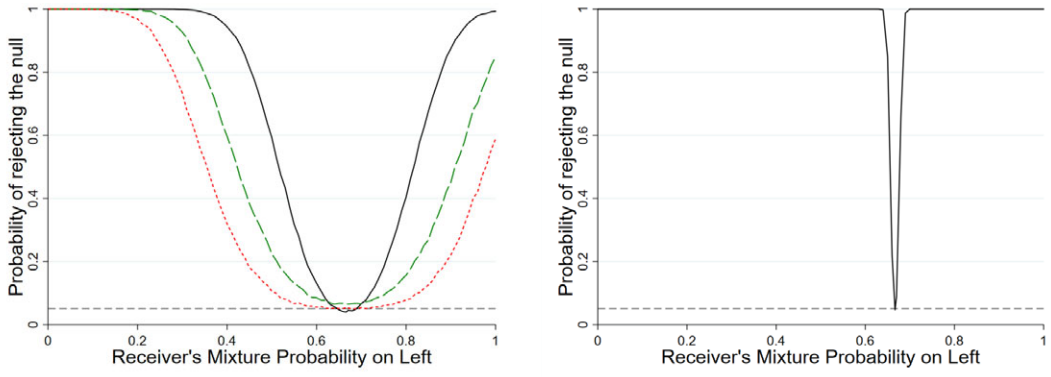
The first column of Table 6 shows the probability of rejecting the null when using the KS test based on the randomized Fisher exact test t-values. Note that the test is valid. Specifically, if the null is true, that is, $\theta = 2/3$, then the null is rejected with probability approximately 0.05. By contrast, if the null is false, for example, $H_a(0.65)$ is true, that is, the server's true winning probability is $p_L(0.65) = 0.6535$ for serves left and $p_R(0.65) = 0.6460$ for serves right, then H_0 is rejected at the 5% level with probability 0.850. The second column of Table 6 shows that the KS test based on the Pearson goodness-of-fit p-values is not valid: it rejects the null hypothesis at the 5% significance level with probability 1 when the null is true. The third column shows that the Pearson joint test (see WW, p. 1527, for a description of this test) is also not valid.

Figure 14(b) shows the power of KS test based on the randomized Fisher exact test t-values for all values of θ. It shows that our test, coupled with a large data set, yields a powerful test of the joint null hypothesis of equality of winning probabilities. The power functions for the Pearson joint test and the KS test based on the p-values from the Pearson goodness-of-fit test are omitted since, as shown in Table 6, neither test is valid.[32]

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

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

[32] As a robustness check, Appendix D in the Online Supplementary Material reproduces the results reported here, but where the simulated data matches the characteristics of the observed data, point game by point game, rather than just in aggregate.



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

FIGURE 14. Power functions for KS test based on t-values (solid), p-values (dotted), and Pearson joint (dashed).

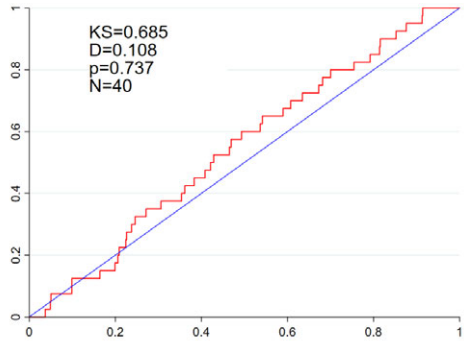
dotted-line power function (the curve at bottom) shows the probability of rejecting H_0 when $H_a(\theta)$ is true for the KS test based on the empirical distribution of the 40 p-values from the Pearson goodness-of-fit test.[33] Importantly, it shows that this test is valid for the WW sample. The dashed-line power function (middle curve) is for the Pearson joint test, and is the analogue of the power function shown in Figure 4 of WW. The solid-line power function (curve at top) is for the KS test based on the empirical distribution of 40 t-values from the randomized Fisher exact test. This last test is, by far, the most powerful. If, for example, $H_a(0.6)$ is true, then the KS test based on the t's rejects H_0 at the 5% significance level with probability 0.131, while the Pearson joint test and the KS test based on the Pearson goodness-of-fit p-values reject H_0 with probability 0.085 and 0.055, respectively.

Reanalysis of prior findings

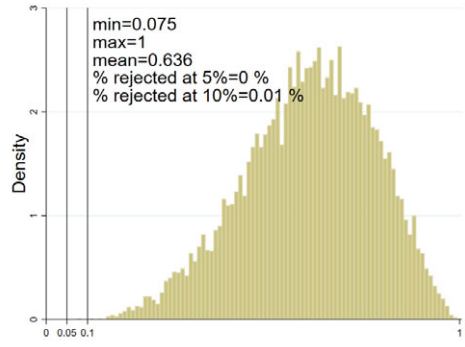
The KS test we propose, based on the randomized Fisher exact test t-values, is valid for all sample sizes and is more powerful than the existing tests used in the literature. Given its greater power, our test has the potential to overturn results in the prior literature based on less-powerful tests.

Using the KS test based on the Pearson goodness-of-fit p-values, WW found that the joint null hypothesis of equality of winning probabilities did not come close to being rejected. Figure 15(a) shows one realization of the empirical distribution of Fisher exact test t-values for the WW data. For this realization, the value of the test statistic is $K = 0.685$ and the associated p-value is 0.737. Figure 15(b) shows that the KS test p-values after 10,000 trials are concentrated around 0.6, and hence are far from the rejection region. The joint null hypothesis of equality of winning probabilities is not rejected even once at the 5% significance level. Thus, the new test confirms the WW finding that

[33]For the power functions reported in Table 6 and Figure 14, the data is simulated 10,000 times for each value of $\theta \in \{0, 0.01, 0.02, \ldots, 0.99, 1\}$ and for $\theta = 2/3$.



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



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

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

the joint null hypothesis of equality of winning probabilities for first serves does not come close to being rejected for male professional tennis players.

HHT studies a data set comprised of ten men's matches, nine women's matches, and eight junior's matches. The men's and women's matches are all from Grand Slam finals, while the juniors matches include the finals, quarterfinals, and second-round matches in both tournaments and Grand Slam matches. HHT found, using the KS test based on Pearson p-values, that the joint null hypothesis of equality of winning probabilities is not rejected for any one of their data sets, or all three jointly. The KS statistics are 0.778 for men (p-value 0.580), 0.577 for women (p-value 0.893), 0.646 for juniors (p-value 0.798), and 0.753 (p-value 0.622) for all 27 matches or 108 point games combined. We show that this conclusion is robust for men, but not for women and juniors, to using the more powerful test based on the t-values.

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

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

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

7. CONCLUSION

We conclude by quantifying the strategic consequences of the failure of female players to perfectly equalize the server's winning probabilities. Recall that in the Hawk-Eye data,

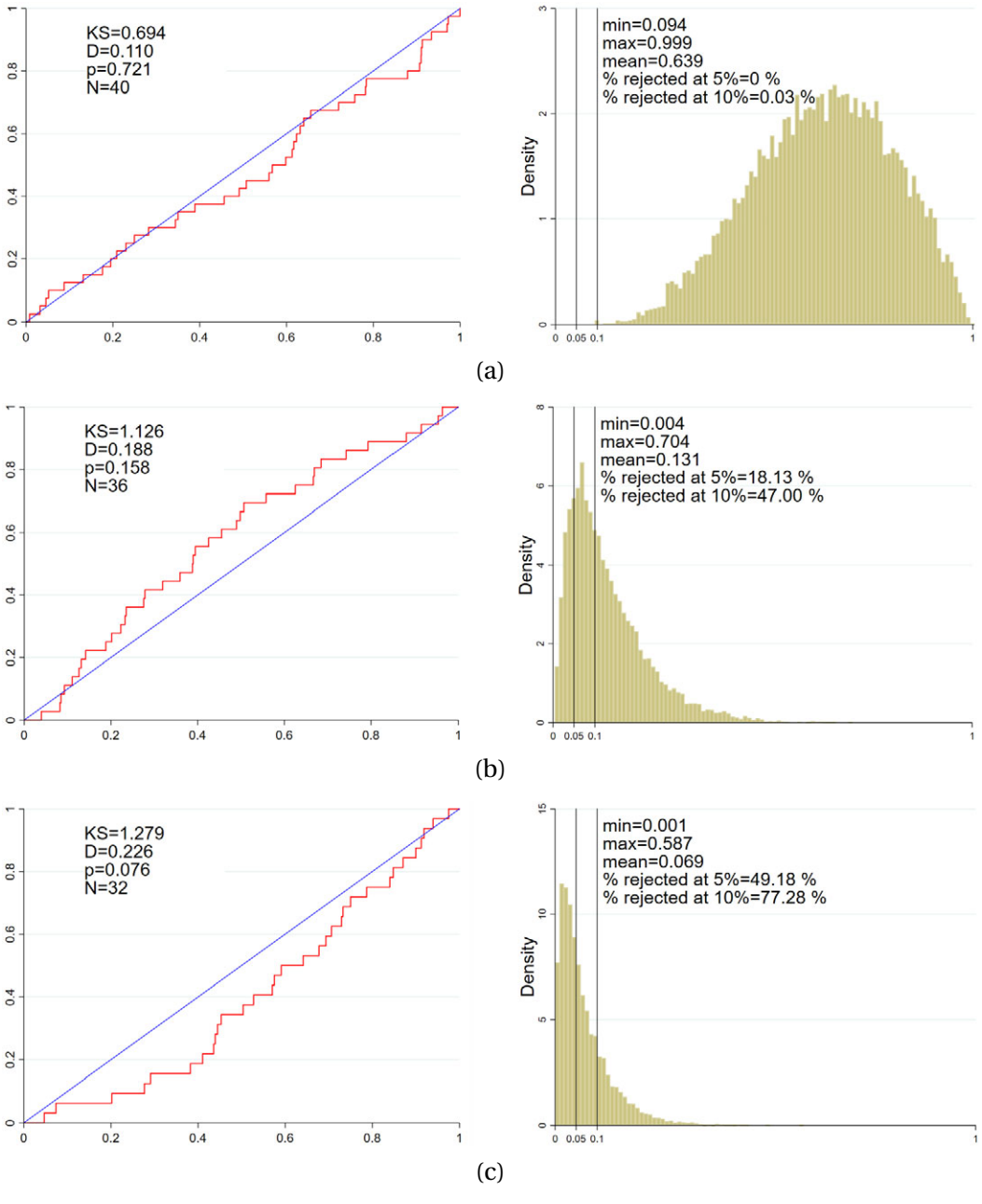


FIGURE 16. (a) KS test for men of $H_0 : p_L^i = p_R^i \ \forall i$ (HHT data). (b) KS test for women of $H_0 : p_L^i = p_R^i \ \forall i$ (HHT data). (c) KS test for Juniors of $H_0 : p_L^i = p_R^i \ \forall i$ (HHT data).

for a sample of 4108 point games of first serves by women, the KS test rejects the joint null hypothesis that winning probabilities are equalized in 44.73% of the trials (see Figure 6). To quantify the strategic consequences, we exploit the illustrative point game in Figure 2. We first identify the deviation from equilibrium play by receivers that is con-

sistent with the 44.73% rejection rate in 4108 point games. We then identify the increase in the probability of winning the match a player can obtain by exploiting the receiver's departure from equilibrium play when serving, while continuing to play equilibrium herself when receiving.

We simulate data for 4108 point games, with 30 serves per point game, under the hypothesis that servers chooses L with equilibrium probability $8/15$, while receivers chooses L with probability θ. Were all receivers to follow the equilibrium mixture, choosing L with probability $\theta = 2/3$, then the probability that the server wins the point is the same for both directions, that is, $p_L(2/3) = p_R(2/3) = 0.65$, and the KS test of the joint null hypothesis that $p_L^i = p_R^i$ for $i \in \{1, \ldots, 4108\}$ is rejected at the 5% level in 5% of the trials. The failure of receivers to equalize winning probabilities implies receivers either underplay or overplay L relative to its equilibrium frequency. Monte Carlo simulations show that $\theta' = 0.6532$ and $\theta'' = 0.6768$ are each consistent with the observed 44.73% rejection rate.

Suppose that all receivers underplay L, following the mixture $\theta' = 0.6532$. The best response for servers is to choose L with probability 1, thereby increasing the probability of winning a point from 0.65 to

$$p_L(0.6532) = 0.58(0.6532) + 0.79(1 - 0.6532) = 0.65283.$$

In a match of two identical players, the probability payoffs in Figure 2 govern the players' payoffs regardless of which player is serving. Consider such a match, where (i) one player deviates from equilibrium, choosing $\theta = 0.6532$ when receiving, and (ii) his opponent exploits the deviating player when serving (choosing L with probability 1) but follows equilibrium when receiving (choosing $\theta = 2/3$). In such a match, the exploiting player wins a point with probability 0.65283 when serving, while the deviating player wins a point with probability 0.65 when serving. One can show that the exploiting player wins increases her probability of winning a three-set match from 0.5 to 0.51393, an increase of 2.786%.[34] This appears to be a strategically and economically significant: obtaining a similar increase in the probability of winning a match by other means could easily require substantially better coaching, substantially more practice, or substantially more raw talent.[35]

This back-of-the-envelope calculation overestimates the strategic consequences of the nonequilibrium play hypothesized. It assumes the server immediately recognizes a receiver's departure from equilibrium play. More important, it assumes that servers can fully exploit (via switching to a pure strategy) receivers who departs from equilibrium play, which is surely unrealistic. In practice, a receiver would change his play if the server adopted a pure strategy. In order to avoid tipping off the receiver, a sophisticated server would exploit receivers by switching to a pure strategy only on the most impor-

[34]The calculation is performed by https://hiddengameoftennis.com/tennis-calculators-markov-win/.

[35]The analogous calculation shows that against a receiver who chooses $\theta'' = 0.6768$, a server can raise her probability of winning a three-set match to 0.51195, a 2.39% increase. The benefit of exploiting a nonequilibrium receiver is even greater in a five-set match, as played in men's tennis.

TABLE A.1. Number of serves and point games after data cleaning.

		Female			Male		
		1st	2nd	N	1st	2nd	N
	All	147,000	57,005	4657	284,109	113,757	7951
(i)	Scoreline	115,014	44,082	4511	230,305	91,341	7690
(ii)	Server?	113,125	43,387	4511	228,802	90,739	7690
(iii)	1st or 2nd?	113,121	42,180	4511	228,785	87,732	7690
(iv)	$\geq$10 serves	110,886	41,376	4108	226,298	86,702	7198

tant points.[36] See https://www.youtube.com/watch?v=ja6HeLB3kwY for Andre Agassi's description of how he exploited Becker by "reading Becker's tongue."

APPENDIX A: DATA CLEANING

There were several steps in the cleaning the data. Table A.1 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.[37] We then eliminated those serves in which there is ambiguity regarding which player is serving (Row (ii)) and those in which there is ambiguity regarding whether the serve is a first or second serve (Row (iii)). Finally, if a point game has fewer than 10 first serves, then we drop the point game and also the associated point game of second serves (Row (iv)).[38]

APPENDIX B: A DETERMINISTIC TEST

Here, we describe and study the properties of a deterministic test, suggested by Hirano (personal communication), of the joint null hypothesis that winning probabilities are equal for serves to the left and serves to the right. Hereafter, we refer to this test as the "Hirano deterministic test." While the test is valid (by construction), we show that it has substantially less power than the KS test based on the randomized Fisher exact test-t values that we develop and employ in the body of the paper. The lower statistical power of the Hirano deterministic test means that it sometimes fails to reject a null hypothesis when the same null is rejected by the KS test based on the randomized Fisher exact test-t values, as we show below.

[36]The importance of a point can be measured as the difference in the probability of winning the match that results from winning rather than losing the current point.

[37]A point can be played more than once, for example, after a set, after a successful challenge of the umpire's decision (on the serve or later in the rally), or after a member of the audience disturbs play. When the same point is played more than once, we use only the serve direction of the first play.

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

The test

We test the joint null hypothesis that $p^i_L = p^i_R$ for each point game $i \in \{1, \ldots, N\}$.[39] Hirano's deterministic test is based on the empirical winning frequencies for serves left and serves right, denoted for point game i by $\hat{p}^i_L = n^i_{SL}/n^i_L$ and $\hat{p}^i_R = n^i_{SR}/n^i_R$, respectively. The Hirano test has three steps. Step I is to compute the test statistic $\hat{X} = \sum_{i=1}^N |\hat{p}^i_L - \hat{p}^i_R|$, which is the sum of the absolute differences of the empirical winning frequencies over N point games. Let $n_S = (n^1_S, \ldots, n^N_S)$, $n_L = (n^1_L, \ldots, n^N_L)$, and $n_R = (n^1_R, \ldots, n^N_R)$ be the marginals for the N point games. Let $F_{|n_S,n_L,n_R}(X)$ denote the distribution of X, the sum of the absolute values of the differences between the left- and right-winning frequencies, conditional on the marginals and under the null hypothesis. Then the two-sided p-value for $\hat{X}$ is

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

Since $F_{|n_S,n_L,n_R}(X)$ is unknown, in Step II we generate B simulated values of the test statistic under the null hypothesis. Recall that in point game i the probability of k winning serves to the left, under the null hypothesis that $p^i_L = p^i_R$ and given the marginals n^i_S, n^i_L, and n^i_R, is given by

$$f\big(k|n^i_S, n^i_L, n^i_R\big) = \frac{\binom{n^i_L}{k}\binom{n^i_R}{n^i_{RS}}}{\binom{n^i_L + n^i_R}{n^i_S}},$$

where $n^i_{RS} = n^i_S - k$. Hence, $f(k|n^i_S, n^i_L, n^i_R)$ is the probability that $\hat{p}^i_L = k/n^i_L$ and $\hat{p}^i_R = (n^i_S - k)/n^i_R$ are the winning frequencies for serves left and serves right. Simulating empirical winning frequencies for each of the N point games according to these distributions, we obtain a simulated value for the test statistic. Let $\hat{X}^*_{|n_S,n_L,n_R}(j)$ denote the jth simulated value of the test statistic. Finally, in Step III, following MacKinnon (2009), the equal-tailed simulated p-value for $\hat{X}$ is[40]

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

Steps I and II were suggested by Hirano, and Step III follows MacKinnon (2009).

Table B.1 shows the test statistics, with the associated simulated p-values and sample sizes, for male and female players and first and second serves, of the test of the null hypothesis that winning probabilities are equalized.

For male players, Hirano's deterministic test reaches the same conclusions as the KS test based on the randomized Fisher exact test t-values: the joint null hypothesis that

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

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

TABLE B.1. Deterministic tests of $H_0 : p_L^i = p_R^i \ \forall i$ for various subsamples, $B = 1000$.

Sample		Men		Women	
		First Serve	Second Serve	First Serve	Second Serve
All	$\hat{X}$	1090.5	1890.7	704.5	1098.9*
	p^*	0.613	0.540	0.252	0.024
	N	7188	6131	4095	3483
With Ranking	$\hat{X}$	1045.9	1813.5	489.0	784.0
	p^*	0.902	0.618	0.493	0.145
	N	6892	5856	2902	2486
Top Receiver	$\hat{X}$	539.2	915.4	244.9	385.9
	p^*	0.240	0.536	0.502	0.362
	N	3462	2926	1461	1254
Nontop Receiver	$\hat{X}$	506.6	898.0	244.1	398.1
	p^*	0.303	0.932	0.766	0.265
	N	3430	2930	1441	1232

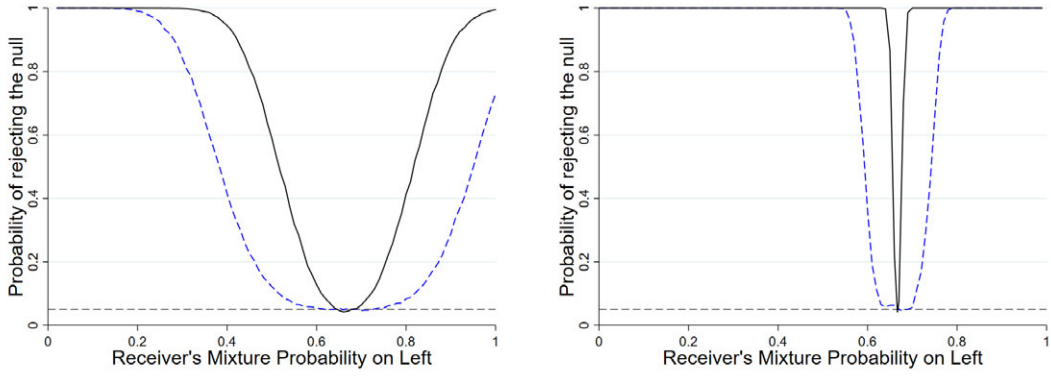
winning probabilities are equalized is not rejected for either the whole sample or any of the subsamples. For female players, Hirano's deterministic test rejects the joint null for second serves, but it does not reject the null for nontop receivers, as did the KS test based on the Fisher exact test t-values.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

The power of the test

We now study the power of Hirano's deterministic test, performing for this test the same simulations reported in Section 6 for the KS test based on the randomized Fisher exact test t-values. Figure B.1 is the analog of Figure 14. The solid lines reproduce the power functions for the KS test based on the randomized Fisher exact test t-values.

The power functions for Hirano's deterministic test (dashed line) are generated as follows. Let $\theta \in [0, 1]$. We first simulate a random sample under the alternative hypothesis $H_a(\theta) : p_L^i = p_L(\theta)$ and $p_R^i = p_R(\theta) \ \forall i \in \{1, \ldots, N\}$. Let $\hat{X} = \sum_{i=1}^{N} |\hat{p}_L^i - \hat{p}_R^i|$ be the associated value of the test statistic, and let n_S, n_L, and n_R be the associated vectors of marginal distributions. We then simulate 1000 values of the test statistic under the null hypothesis that $p_L^i = p_R^i \ \forall i \in \{1, \ldots, N\}$. Let $\hat{F}^*_{|n_S,n_L,n_R}(X)$ be the empirical *c.d.f.* of the simulated test statistic. The null hypothesis is rejected at the 5% significance level if either $\hat{F}^*_{|n_S,n_L,n_R}(\hat{X}) \leq 0.025$ or $1 - \hat{F}^*_{|n_S,n_L,n_R}(\hat{X}) \leq 0.025$, that is, if the realized value of the test statistic for the simulated sample is in the tails of the *c.d.f.* of the simulated test statistic. The power function for the deterministic test is the probability that the null is rejected when $H_a(\theta)$ is true.[41]

[41]For the power functions reported in Figure B.1 and Table B.2, when $N = 40$ and when $N = 7000$, the data is simulated 10,000 times and 1000 times, respectively, for each value of $\theta \in \{0, 0.01, 0.02, \ldots, 0.99, 1\}$ and for $\theta = 2/3$.



$N = 40$ $N = 7000$

FIGURE B.1. Power functions for the deterministic test and for the KS test based on t-values.

It is evident from Figure B.1 that Hirano's deterministic test has substantially less power, for both small and large samples. These results demonstrate the usefulness of the test we develop in the body of the paper. Given the lower power of Hirano's deterministic test, unsurprisingly it fails to reject the null for some subsamples for which the KS test based on the randomized t-values does reject.

Table B.2 shows that Hirano's deterministic test is especially low powered, in comparison to our KS test based on the randomized Fisher exact test t-values, in the neighborhood of the null hypothesis.

APPENDIX C

This Appendix has three parts. The first part proves that when the null hypothesis $p_L^i = p_R^i$ is true for point game i, then the randomized Fisher Exact test value t^i is distributed $U[0, 1]$. The second part shows that the randomized Fisher exact test t^i's continues to be distributed $U[0, 1]$ even if there is serial correlation in the direction of the serve. Finally, the third part demonstrates that the KS test, based on the p-values from the Pearson goodness-of-fit test, will fail when the number of point games grows large while the number of serves per point game is held fixed.

TABLE B.2. Rejection rate for H_0 at the 5% level, $N = 7000$.

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

Proof that $t^i \sim U[0,1]$ under the null

Assume that $p_L^i = p_R^i = p$. As shown by Fisher (1935), the random variable n_{LS}^i, with support $\{\underline{n}_{LS}^i, \underline{n}_{LS}^i + 1, \ldots, n_L^i\}$ and where $\underline{n}_{LS}^i = \max\{n_S^i - n_R^i, 0\}$, has *p.d.f.*,

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

and *c.d.f.*,

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

Let t^i be the random test statistic defined, as described in Section 4, by

$$t^i \sim \begin{cases} U\left[0, F\left(n_{LS}^i | n_S^i, n_L^i, n_R^i\right)\right] & \text{if } n_{LS}^i = n_S^i - n_R^i, \\ U\left(F\left(n_{LS}^i - 1 | n_S^i, n_L^i, n_R^i\right), F\left(n_{LS}^i | n_S^i, n_L^i, n_R^i\right)\right] & \text{otherwise.} \end{cases}$$

We prove that t^i is distributed $U[0,1]$.

The following claim holds for any discrete random variable, and thus it holds in our context as well.

Claim. *Let X be a discrete random variable with support $\{x_1, \ldots, x_K\}$, where $x_1 < x_2 < \cdots < x_K$, with p.d.f. $f(x_k) > 0$ and associated c.d.f. $F(x_k)$. Let t be the random variable defined by*

$$t \sim \begin{cases} U\left[0, F(x_1)\right] & \text{if } X = x_1, \\ U\left(F(x_{k-1}), F(x_k)\right] & \text{if } X = x_k. \end{cases}$$

Then t is distributed $U[0,1]$.

Proof. First, note that the support of t is the interval $[0,1]$ since the union of the intervals $[0, F(x_1)]$, $(F(x_1), F(x_2)], \ldots, (F(x_{K-1}), F(x_K)]$ is $[0,1]$. Furthermore, since these intervals are disjoint, then any $z \in [0,1]$ is an element of exactly one interval.

Let $z \in [0,1]$ be arbitrary. We need to show that $\Pr\{t \le z\} = z$. If $z \le F(x_1)$, then

$$\Pr\{t \le z\} = \Pr\{0 \le t \le z | X = x_1\} f(x_1) = \frac{z}{F(x_1)} f(x_1) = z.$$

If $z > F(x_1)$, then there is a unique $k' > 1$ such that $z \in (F(x_{k'-1}), F(x_{k'})]$. Then

$$\begin{aligned} \Pr[t \le z] &= F(x_{k'-1}) + \Pr\{x_{k'-1} < t \le z\} \\ &= F(x_{k'-1}) + \Pr\{t \le z | X = x_{k'}\} f(x_{k'}) \end{aligned}$$

$$= F(x_{k'-1}) + \frac{z - F(x_{k'-1})}{F(x_{k'}) - F(x_{k'-1})} f(x_{k'})$$
$$= z.$$

This completes the proof. □

The proof requires that t be drawn from $U(F(x_{k-1}), F(x_k)]$ with probability $f(x_k)$. In our context, the *c.d.f.* of interest is $F(n^i_{LS}|n^i_S, n^i_L, n^i_R)$. If the null hypothesis that $p^i_L = p^i_R = p$ is *not* true, then the probability t^i is drawn from

$$U\big(F\big(n^i_{LS} - 1|n^i_S, n^i_L, n^i_R\big), F\big(n^i_{LS}|n^i_S, n^i_L, n^i_R\big)\big]$$

need not be $f(n^i_{LS}|n^i_S, n^i_L, n^i_R)$ as n^i_{LS} need not be distributed according to Fisher's formula. The t^i value might, for example, be drawn from the interval $[0, \underline{n}^i_{LS}]$ with probability greater than $f(\underline{n}^i_{LS}|n^i_S, n^i_L, n^i_R)$.

Serial correlation and the randomized Fisher exact test

We show that the t-values in the randomized Fisher exact test remain uniformly distributed even when the server exhibits serial correlation in the direction of the serve. Hence, our test of the null that $p^i_L = p^i_R \ \forall i$ is not undermined by the presence of serial correlation in the direction of the serve that we find in our data.

We need to show that the formula for the Fisher exact test

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

continues to hold if the server exhibits serial correlation in the direction of serve.[42] Assume that the number of winning serves to the left, conditional on the number of left serves, is independent of the number of winning serves to the right, conditional on the number of right serves, that is,

$$P(n_{LS}, n_{RS}|n_L, n_R) = P(n_{LS}|n_L)P(n_{RS}|n_R).$$

A simple example of a DGP that would satisfy this assumption is that the server alternates between serves left and serves right.[43] More generally, the assumption is satisfied if the probability of a serve left or right depends only on the history of directions of past serves.

[42]For notational convenience, we suppress the index i for the point game.

[43]This, of course, would be inconsistent with equilibrium play.

We now show that Fisher's formula holds given the assumption above. We have that

$$\begin{aligned} f(n_{LS}|n_S, n_L, n_R) &= \frac{P(n_{LS}, n_{RS}, n_L, n_R)}{P(n_S, n_L, n_R)} \\ &= \frac{P(n_{LS}, n_{RS}, n_L, n_R)}{P(n_S, n_L, n_R)} \frac{P(n_L, n_R)}{P(n_L, n_R)} \\ &= \frac{P(n_{LS}, n_{RS}|n_L, n_R)}{P(n_S|n_L, n_R)}. \end{aligned}$$

By the conditional independence assumption, we have

$$\begin{aligned} P(n_{LS}, n_{RS}|n_L, n_R) &= P(n_{LS}|n_L)P(n_{RS}|n_R) \\ &= \binom{n_L}{n_{LS}} p^{n_{LS}}(1-p)^{n_L - n_{LS}} \binom{n_L}{n_{RS}} p^{n_{RS}}(1-p)^{n_R - n_{RS}}, \end{aligned}$$

where the second equality follows from the null hypothesis that $p_L = p_R = p$. Likewise, we have

$$P(n_S|n_L, n_R) = \binom{n_L + n_R}{n_S} p^{n_S}(1-p)^{n_L + n_R - n_S}.$$

Thus,

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

which completes the proof.

Failure of the KS test based on the Pearson p-values

We show that WW's test of the equality of winning probabilities, that is, the KS test based on the p-values from the Pearson goodness-of-fit test, fails when the number of point games grows large, while at the same time, the number of serves per point game is fixed. It fails as, under the null hypothesis, the empirical *c.d.f.* of p-values converges to a discrete distribution rather than the $U[0, 1]$ distribution.

We first prove that the test fails in the simpler setting of testing that each coin in a collection of N coins is fair. In this discussion, a coin is the analogue of a point game and a toss of the coin is the analogue of a serve in the point game. We consider coins rather than point games since the probability of a Heads is $p(H) = 0.5$ under the null hypothesis, whereas the probability of winning a point on a serve is unknown. This simplifies the explanation of why the KS test based on the Pearson p-values is not valid when N, the number of coins (or point games) grows large, while the number of coin tosses (or serves) is fixed.

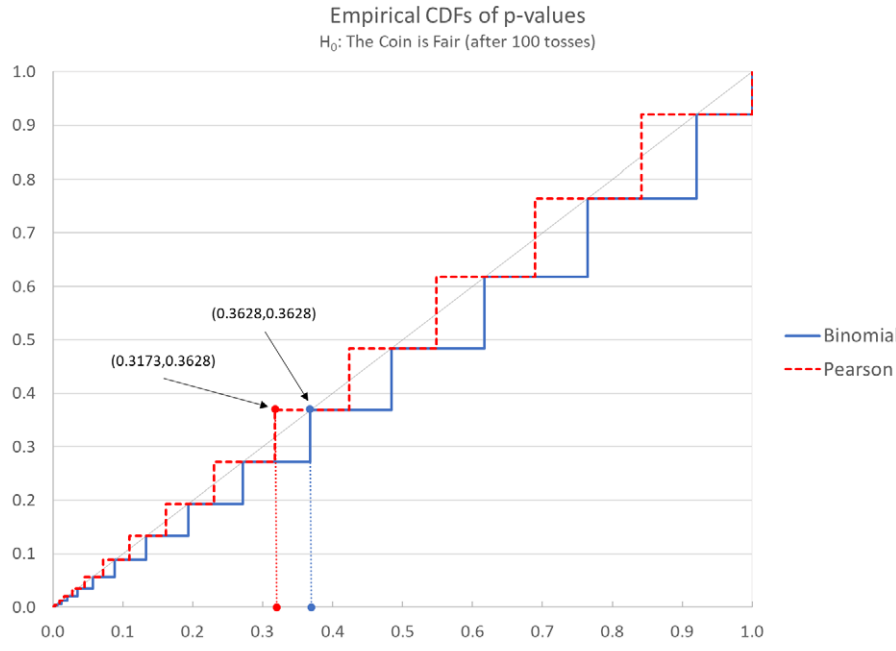


FIGURE C.1. Theoretical *c.d.f.*'s of binomial and Pearson p-values for $H_0 : p_H = 0.5$.

The theoretical *c.d.f.* of the binomial p-values for the null hypothesis that a single coin tossed $n = 100$ times is fair is the solid line in Figure C.1.[44] The theoretical distribution of the Pearson goodness-of-fit p-values when testing the null hypothesis, after 100 tosses, that the coin is fair is the dashed line and is constructed as follows: Let n_H and n_T denote the number of heads and tails, respectively, after n tosses of the coin. The test statistic is

$$Q^n = \frac{(n_H - n/2)^2}{n/2} + \frac{(n_T - n/2)^2}{n/2},$$

where $n/2$ is the expected number of each outcome after n tosses. Under the null hypothesis, Q^n is asymptotically distributed chi-square with 1 degree of freedom ($\chi^2(1)$) as n grows large. Let $F(x; 1)$ denote the *c.d.f.* of the $\chi^2(1)$ distribution. If $\hat{q}^n$ is the realized value of Q^n, then the associated p-value is $1 - F(\hat{q}^n; 1)$. If $n_H = 45$, for example, then the binomial p-value is 0.3628 and Pearson p-value is 0.3173 (since $\hat{q}^{100} = 1.0$). Figure C.1 illustrates that the probability of a Pearson p-value of 0.3173 or less is 0.3682.

Both *c.d.f.*'s in Figure C.1 approach the $U[0, 1]$ distribution as n grows large, but both are discrete for finite n.

Suppose next that we have many independent coins and for each coin i we obtain a Pearson p-value p_i after $n = 100$ tosses. As the number N of coins approaches infinity,

[44]Recall that a p-value is the probability of obtaining a test result at least as extreme as the result actually observed, when the null hypothesis is true. The theoretical *c.d.f.* of the binomial p-values is exact: for each possible realized value x of a p-value, the probability of a p-value less than or equal to x is itself x. This means that the "steps" on the *c.d.f.* of binomial p-values fall on the 45-degree line.

by the Glivenko–Cantelli theorem the empirical distribution of N Pearson p-values approaches the (discrete) theoretical distribution of Pearson p-values shown above. The maximal distance between the theoretical *c.d.f.* of the Pearson p-values and the *c.d.f.* of the uniform distribution approaches 0.0796.[45] Thus, as N approaches infinity the maximal distance between the empirical *c.d.f.* of Pearson p-values and the uniform *c.d.f.* approaches 0.0796, and the KS test statistic approaches $K = 0.0796\sqrt{N}$. By Mood, Graybill, and Boes (1974, Theorem 1, p. 508), the limiting distribution of the KS test statistic is

$$H(x) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2} \quad \text{for } x > 0.$$

Hence, $H(0.0796\sqrt{N})$ approaches one as N approaches infinity. The null hypothesis that each of N coins is fair is rejected at any significance level, even though the null is true.

Suppose next that we have a pair of coins, a "Left" coin and a "Right" coin, and the coins have probabilities $p_L(H)$ and $p_R(H)$, respectively, of coming up heads when tossed. Testing the hypothesis that $p_L(H) = p_R(H) = p(H)$, after n_L tosses of the left coin and n_R tosses of the right coin, is exactly analogous to testing the hypothesis that $p_L^i = p_R^i = p^i$, that is, winning probabilities are equalized, in a single point game i, after n_L^i and n_R^i serves left and right. As in the example above, the Pearson p-value has a discrete distribution (which depends on the true, but unknown, value of $p(H)$) under the null hypothesis.[46] As the number of pairs of coins approached infinity, the empirical distribution of Pearson p-values will approach its theoretical distribution. The KS test will reject the null hypothesis that $p_L(H) = p_R(H)$ for every pair of coins in the collection.

The KS test based on the randomized Fisher exact t-values, by contrast, is a valid test for any number N of coins and tosses n (or point games and serves). This follows since, under the relevant null hypothesis—the coins are fair or winning probabilities are equal for serves in each direction—the randomized Fisher exact t-values are exactly and continuously distributed $U[0, 1]$, as established at beginning of this Appendix. As N grows large, the empirical *c.d.f.* approaches the $U[0, 1]$ distribution.

In both examples above, the collections considered were homogeneous under the null: in the first example, each coin in the collection was fair, while in the second example each pair of coins had the same probability $p(H)$ of heads. In the Hawk-Eye data, the point games are heterogenous, with p^i varying across i, and hence the theoretical distribution of the Pearson p-value will vary with i as well. Nonetheless, there is no reason to expect under the null hypothesis $p_L^i = p_R^i \ \forall i$ that the empirical distribution of Pearson p-values will approach the $U[0, 1]$ distribution as the number of point games grows large.

The Monte Carlo simulations reported in Appendix C study the behavior of the KS test when point games are heterogenous, with winning probabilities that match those of the Hawk-Eye data. The results reported there show that when there are 7000 point

[45]The maximum distance is at the first "jump down" of the Pearson *c.d.f.* on the right- hand side.

[46]See Section 4 (subsection on Comparing Tests) for a description of the Pearson test.

games, the KS test based on the Pearson p-values rejects the null hypothesis that winning probabilities are equalized for serves in different directions, even when the null is true.

References

Camerer, Colin (2003), *Behavioral Game Theory Experiments in Strategic Interaction*. Princeton University Press. [982]

Chiappori, Pierre, Steven Levitt, and Timothy Groseclose (2002), "Testing mixed strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer." *American Economic Review*, 92, 1138–1151. [982]

Cooper, David, John Kagel, Wei Lo, 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. [985]

Fisher, Ronald (1935), *The Design of Experiments*. Hafner Publishing Company, New York. [991, 1014]

Gauriot, Romain, Lionel Page, and John Wooders (2023), "Supplement to 'Expertise, gender, and equilibrium play'." *Quantitative Economics Supplemental Material*, 14, https://doi.org/10.3982/QE1563. [990]

Gonzalez-Diaz, Juan, Olivier Gossner, and Brian Rogers (2012), "Performing best when it matters most: Evidence from professional tennis." *Journal of Economic Behavior & Organization*, 84, 767–781. [986]

Hsu, Shih-Hsun, Chen-Ying Huang, and Cheng-Tao Tang (2007), "Minimax play at Wimbledon: Comment." *American Economic Review*, 97, 517–523. [984]

Klaassen, Franc and Jan R. 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. [1003]

Klaassen, Franc and Jan R. Magnus (2014), *Analyzing Wimbledon: The Power of Statistics*. Oxford University Press. [1004]

Kocher, Martin, Marc Lenz, and Matthias Sutter (2012), "Psychological pressure in competitive environments: New evidence from randomized natural experiments." *Management Science*, 58, 1585–1591. [986]

Kovash, Kenneth and Steven Levitt (2009), "Professionals do not play minimax: Evidence from major league baseball and the national football league." NBER working paper 15347. [982]

Lehmann, Erich and Joseph Romano (2005), *Testing Statistical Hypotheses*. Springer, New York. [992]

Levitt, Steven, John List, and David Reiley (2010), "What happens in the field stays in the field: Professionals do not play minimax in laboratory experiments." *Econometrica*, 78, 1413–1434. [985]

Levitt, Steven, John List, and Sally Sadoff (2011), "Checkmate: Exploring backward induction among chess players." *American Economic Review*, 101, 975–990. [985]

MacKinnon, James (2009), "Bootstrap hypothesis testing." In *Handbook of Computational Econometrics* (David Belsey and Erricos John Kontoghiorghes, eds.), 183–210. John Wiley & Sons. [1011]

Mood, Alexander, Franklin Graybill, and Duane Boes (1974), *Introduction to the Theory of Statistics*. McGraw Hill, New York. [993, 1018]

O'Neill, Barry (1987), "Nonmetric test of the minimax theory of two-person zero-sum games." *Proceedings of the National Academy of Sciences*, 84, 2106–2109. [983, 996, 997]

Palacios-Huerta, Ignacio (2003), "Professionals play minimax." *Review of Economic Studies*, 70, 395–415. [982]

Palacios-Huerta, Ignacio and Oscar Volij (2008), "Experientia docent: Professionals play minimax in laboratory experiments." *Econometrica*, 76, 71–115. [985]

Paserman, Daniele (2010), "Gender differences in performance in competitive environments: Evidence from professional tennis players." Report. [986]

Rothenberg, Ben (2017), "Filling a weak spot in women's tennis: The serve." *New York Times*, https://www.nytimes.com/2017/09/03/sports/tennis/us-open-wta-tour-serving.html. [984]

Tocher, Keith (1950), "Extension of the Neyman–Pearson theory of tests to discontinuous variates." *Biometrika*, 37, 130–144. [992]

Van Essen, Matthew and John Wooders (2015), "Blind stealing: Experience and expertise in a mixed-strategy poker experiment." *Games and Economic Behavior*, 91, 186–206. [985]

Walker, Mark and John Wooders (2001), "Minimax play at Wimbledon." *American Economic Review*, 91, 1521–1538. [982]

Walker, Mark, John Wooders, and Rabah Amir (2011), "Equilibrium play in matches: Binary Markov games." *Games and Economic Behavior*, 71, 487–502. [987]

Wooders, John (2010), "Does experience teach? Professionals and minimax play in the lab." *Econometrica*, 78, 1143–1154. [985]

Co-editor Peter Arcidiacono handled this manuscript.

Manuscript received 2 March, 2020; final version accepted 30 January, 2023; available online 14 March, 2023.