Minimax Play at Wimbledon

By Mark Walker and John Wooders*

In many strategic situations it is important that one's actions not be predictable by one's opponent, or by one's opponents. Indeed, the origins of modern game theory lie in the attempt to understand such situations. The theory of mixed-strategy play, including von Neumann's Minimax Theorem and the more general notion of a Nash equilibrium in mixed strategies, remains the cornerstone of our theoretical understanding of strategic situations that require unpredictability.

Many experiments designed to test the theory of mixed-strategy play using human subjects have been carried out over the past 40 years or more. The theory has not fared well.[1] The theory's consistent failure in experimental tests raises the question whether there are *any* strategic situations in which people behave as the theory predicts.

We develop a test of the minimax hypothesis using field data from championship professional tennis matches, and we find that win rates in the serve and return play of top professional tennis players are consistent with the minimax hypothesis. However, the players' choices are not consistent with the serial independence implied by the minimax hypothesis: even the best tennis players tend to switch from one action to another too often.

When we apply the same statistical tests to experimental data, both the equilibrium mixing proportions and serial independence of choices are soundly rejected. Our results therefore provide some evidence that play by highly motivated and highly experienced players may conform more closely to the theory of mixed-strategy equilibrium than the play that has been observed in experiments.

We begin from the observation that games are not easy to play, or at least to play well. This is especially true of games requiring unpredictable play. Consider poker—say, five-card draw poker. The rules are so simple that they can be learned in a few minutes' time. Nevertheless, a player who knows the rules and the mechanics of the game but has little experience actually *playing* poker will not play well.[2] Similarly, in experiments on minimax play the rules of the game have typically been simple, indeed transparently easy to understand. But subjects who have no experience actually *playing* the game are not likely to understand the game's strategic subtleties—they are likely to understand how to *play* the game, but not how to play the game *well.* Indeed, it may simply not be possible in the limited time frame of an experiment to become very skilled at playing a game that requires one to be unpredictable.

Professional sports, on the other hand, provide us with strategic competition in which the participants have devoted their lives to becoming experts at their games, and in which they are often very highly motivated as well. Moreover, situations that call for unpredictable play are nearly ubiquitous in sports: The pitcher who "tips" his pitches is usually hit hard, and batters who are known to "sit on" one pitch usually don't last long. Tennis players must mix their serves to the receiver's forehand and backhand sides; if the receiver knew where the serve was coming, his returns would be far more effective. Point guards who can only go to their right don't make it in the NBA. Thus, while the players' recognition of the "correct" way to mix

* Department of Economics, University of Arizona, Tucson, AZ 85721 (e-mail: mwalker@arizona.edu; jwooders@bpa.arizona.edu). For their helpful comments we are indebted to Robert J. Aumann, William Horrace, Thomas R. Palfrey, Robert W. Rosenthal, Jason M. Shachat, and Vernon L. Smith, and to an anonymous referee. Ed Agranowitz provided indispensable assistance in obtaining data, and Barry O'Neill graciously provided us with the data from his experiment. William Janss and Zeynep Kocabiyik Hansen provided valuable research assistance. This research was partially supported by Grant No. SBR–9710289 from the National Science Foundation.

[1] See, for example, Figure 1 in Ido Erev and Alvin E. Roth (1998), and their accompanying discussion, which describes 12 such experiments.

[2] The reader can verify this proposition by buying some chips and sitting down at a table at Binion's in Las Vegas.

in these situations may be only subconscious, any significant deviation from the correct mixture will generally be pounced upon by a sophisticated opponent.[3]

As empirical tests of the minimax hypothesis, however, sports are generally inferior to experiments. In the classic confrontation between pitcher and batter, for example, there are many actions available (fastball, curve, change-up, inside, outside, etc.), and the possible outcomes are even more numerous (strike, ball, single, home run, fly ball, double-play grounder, etc.). Difficulties clearly arise in modeling such situations theoretically, in observing players' actions, and in obtaining sufficient data to conduct informative statistical tests.

Tennis, however, provides a workable empirical example: although speed and spin on the serve are important choices, virtually every (first) serve is delivered as far toward one side or the other of the service court as the server feels is prudent, and the serve is for many players an extremely important factor in determining the winner of the point. Moreover, theoretical modeling is tractable (each point has only two possible outcomes: either the server wins the point, or the receiver does); the server's actions are observable (it is easy to see whether he has served to the receiver's forehand or backhand); and data is relatively plentiful (long matches contain several hundred points played by the same two players).

Following this idea, we use simple 2 × 2 games as a theoretical model of the serve and its relation to the winning of points in a tennis match. We have constructed a data set that contains detailed information about every point played in ten professional tennis matches. Each match provides us with four 2 × 2 "point games" with which to test the minimax hypothesis, giving us a total of 40 point games. In each of the 40 point games we use the server's "win rates"—the observed relative frequencies with which he won points when serving to the receiver's left or to his right—to test whether his winning probabilities are indeed the same for both serving directions, as the theory says they should be. In only one of the 40 point games is minimax play rejected at the 5-percent level. This rejection rate is actually slightly below the rate predicted by the random character of equilibrium play.

In addition to equality of players' winning probabilities, equilibrium play also requires that each player's choices be independent draws from a random process. We conduct tests of randomness, and find that the tennis players switch their serves from left to right and vice versa too often to be consistent with random play. This is consistent with extensive experimental research in psychology which indicates that people who are attempting to behave truly randomly tend to "switch too often." The same tests reveal far greater deviation from randomness in experimental data.

I. A Model of the Serve in Tennis

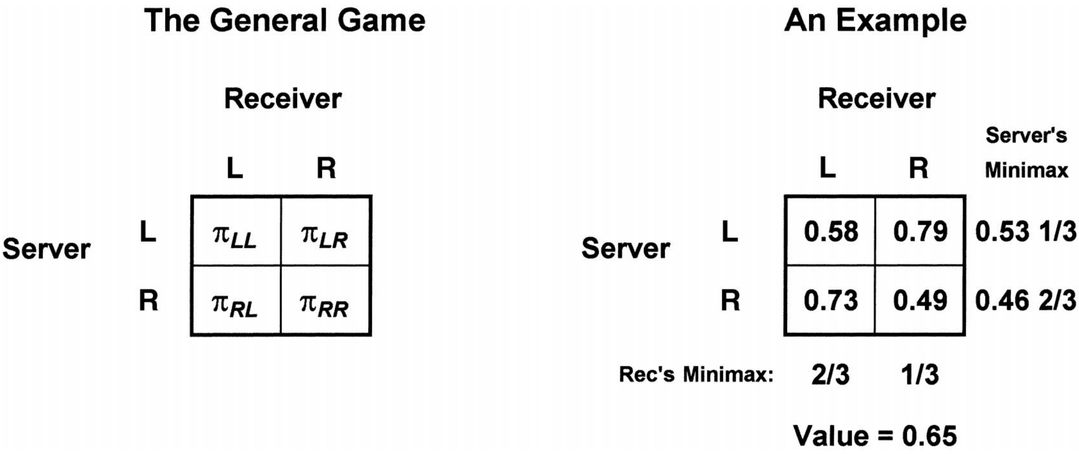
We model each point in a tennis match as a simple 2 × 2 normal-form game between two players.[4] A typical such *point game* is depicted in Figure 1. Each point in a tennis match is begun by one of the players placing the ball in play, or "serving." We assume that the two actions available to the server are to serve either to the receiver's left (L) or to the receiver's right (R). Simultaneously with the server's decision, the receiver is assumed to guess whether the serve will be to the left or to the right—i.e., he makes a decision, perhaps only subconsciously, to "overplay" to one side or the other.[5]

After the server and the receiver have both made their left-or-right choices for the serve, the winner of the point is determined—perhaps immediately (if the serve is not returned successfully), or perhaps after many subsequent

[3] After a recent match, Venus Williams said she had shown her opponent, Monica Seles, several different types of serves. "You have to work on that, because it's very easy to become one-dimensional and just serve to your favorite space and the person is just waiting there." Seles responded "She mixed it up very well . . . I really love that part of her game."

[4] Essentially the same 2 × 2 model appears in Avinash Dixit and Barry Nalebuff (1991).

[5] The point game can be modeled differently. For example, the server can be given more choices (serve to the body; use a flat, slice, or kick serve), and the receiver, instead of "guessing," can choose a location across the baseline where he can position himself to await the serve. These alternative models of the point game make the same predictions as our 2 × 2 point-game model.



The General Game

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

An Example

Server \ Receiver	L	R	Server's Minimax
L	0.58	0.79	0.53 1/3
R	0.73	0.49	0.46 2/3
Rec's Minimax:	2/3	1/3	

Value = 0.65

FIGURE 1. THE POINT GAME

Note: Outcomes (cell entries) are the probability the Server wins the point.

strokes by each player, or perhaps even after a second serve is played (if the first serve turns out to be a fault). We do not attempt to model the play after the serve, but instead adopt a reduced-form representation of it: each player's payoffs in the four cells of the game matrix are the respective probabilities that he will ultimately win the point, conditional on the left-or-right choices each of the players has made on the serve.

The server's probabilities of winning the point are denoted by the four numbers π_{sr}, where s is the server's choice (L or R) and r is the receiver's choice (L or R). Since one player or the other must win the point, the receiver's probabilities of winning are the numbers $1 - \pi_{sr}$. We assume that each player cares only about winning the point; therefore the winning probabilities π_{sr} and $1 - \pi_{sr}$ are the players' payoffs in the 2×2 point game.[6] Because the game is constant sum, it is completely determined by the server's probabilities π_{sr}, as in Figure 1. (Figure 1 includes a numerical example. The example's payoff numbers π_{sr} are hypothetical, but capture salient features of the data.)

We assume that every point game we will encounter satisfies the inequalities $\pi_{LL} < \pi_{RL}$ and $\pi_{RR} < \pi_{LR}$ (i.e., the server is more likely to win the point if he serves away from the direction the receiver is overplaying) as well as the inequalities $\pi_{LL} < \pi_{LR}$ and $\pi_{RR} < \pi_{RL}$ (the server is less likely to win the point if the receiver overplays in the direction the server has chosen). This is equivalent to the following assumption:

ASSUMPTION 1: *Every point in a tennis match is played as a 2×2 constant-sum normal-form game with a unique equilibrium in strictly mixed strategies.*

Both our theoretical and our empirical analysis would be simpler if every point game in every tennis match were the same—i.e., if there were no variation in the four probability payoffs π_{sr} over the course of a match or across matches. This is highly unlikely, however. The probability payoffs in a point game clearly depend upon the abilities of the specific two people who are playing the roles of server and receiver. The probabilities will therefore vary in matches between different people, and perhaps even across matches involving the same pair of opponents but played on different surfaces or under different weather conditions. Moreover, the probabilities will typically vary even within a single match, because the serve alternates between the two players in successive games.

[6] The tennis *match* consists of repeated play of point games. We address below the relation between the point games and a player's strategy for the match.

Further, even when holding the server and receiver fixed, as is done within a single game, the points that make up the game alternate between "deuce-court" points and "ad-court" points. Because of the players' particular abilities, the probability payoffs for a deuce-court point will generally differ from the probabilities for an ad-court point.

In a given match, then, there are typically at least four distinct point games, identified by which player has the serve and by whether it is a deuce-court point or an ad-court point. We assume that there is no further variability in the point games within a single match.

ASSUMPTION 2: *There are four point games in a tennis match, distinguished by which player is serving for the point, and by whether the point is a deuce-court point or an ad-court point.*

The Tennis Match as a Game. A player in a tennis match is presumably interested in winning points only as a means to his ultimate goal of winning the match. The fact that the point games are merely the elements of a larger (infinite-horizon, extensive-form) game raises an immediate question: is it appropriate to assume, as we are doing, that the players' payoffs in the point game are the probabilities they will win the point? The link between the point games and the "match game" is provided by the main result in Walker and Wooders (2000), where a class of games called *binary Markov games* is defined and analyzed. They show that equilibrium play in such games (of which tennis is an example) requires that a player play each point as if it were the only point: his play should be independent of the score (except to the extent that it directly affects the probability payoffs π_{sr}), and independent of the actions or outcomes on all previous points.[7]

II. On Testing the Theory

Our simple theoretical model of tennis, when combined with the equilibrium result from Walker and Wooders (2000), makes some predictions about tennis players' behavior that we can subject to empirical testing. The theory's first implication is that for every point of a tennis match each of the players will make his left-or-right choice according to his minimax mixture for the associated point game. The observed choices in a given match will therefore be independent draws from a binomial process which depends upon (a) which player is serving and (b) whether the point is a deuce-court point or an ad-court point; and the binomial process is otherwise independently and identically distributed (i.i.d.) across all serves in the match. Furthermore, if the four probability payoffs π_{sr} in a point game are known, then it is straightforward to calculate each player's equilibrium mixture. It would seem to be straightforward, then, to simply test whether the observed frequencies of a player's left and right choices [separated according to (a) and (b)] could have been the result of his equilibrium i.i.d. binomial mixture process, in just the same way that tests of the minimax hypothesis have been performed with experimental data.

However, in a tennis match the entries π_{sr} in the payoff matrix are not known, nor can we observe the receiver's choices, and therefore we cannot estimate the numbers π_{sr}. The only elements of the point game that are observable in an actual tennis match are (1) the server's action on each first serve (was the serve to the left or to the right?), and (2) which player ultimately won the point. In a given point game, if the players are playing according to the equilibrium, which is in mixed strategies, then each player's expected payoff from playing left must be the same as his expected payoff from playing right—i.e., a player must have the same probability of winning the point, whichever direction he serves, and his observed win rates can be used to test that hypothesis.

III. The Data

Our data set was obtained from videotapes of ten tennis matches between highly ranked professional players in the four so-called major, or Grand Slam, tournaments and the year-end Masters tournament. All but two of the matches were the final (championship) match of the respective tournament. There were several criteria

[7] Martina Navratilova has said that on the night before she was to play in the 1990 Wimbledon final she condensed her strategy to just a few words: "I had to keep my mind off winning: . . . Think about that point and that point only." (John Feinstein, 1991.)

that we required the matches to satisfy for inclusion in our data set: that winning the match be important to both players (hence the Grand Slam and Masters tournaments); that the players be well known to one another, so that each would enter the match with a good sense of the probability payoffs π_{sr}; and that the matches be long enough to contain many points,[8] in order to have enough observations to make our statistical tests informative—specifically, so that the tests would be likely to reject the minimax hypothesis in cases where it is false (in other words, so that the tests would have adequate power).

Recall that every tennis match contains four point games, so in our ten matches we have data for 40 point games in all. Note that in Table 1, where the data are summarized, the matches are separated by horizontal lines, and there are four rows for each match. Each row corresponds to a point game. Indeed, it will be helpful to think of each row of Table 1 as the data from an "experiment," for which we model the data generating process as a 2 × 2 point game, as in Section I. We will want to test whether the data in these experiments could have been generated by *equilibrium* play of the relevant point game.

The data set contains the following information for every point in every one of the ten matches: the direction of the point's first serve (left, center, or right), and whether or not the server ultimately won the point. The data are presented in Table 1 (with serves to the center omitted: only 6 percent of first serves were to the center, so they would have a negligible effect on our results). The columns labeled Serves in Table 1 indicate, for each match, server, and court (i.e., for each "experiment"), the number of times the direction of the first serve was left (L) or right (R). The columns labeled Points Won indicate, for each direction of first serve, the number of times the server ultimately won the point.[9] The relative frequency of each direction of first serve (the observed mixture) is given in the Mixture columns, and the relative frequencies with which points were won (the "win rate") for each direction are given in the Win Rates columns. The winner of the match is indicated in boldface.

In our data set the players had on average 160 first serves but only 63 second serves. Since the number of second serves from either court is generally small (averaging just 33 from the deuce court and 30 from the ad court in our matches), we analyze only first serves.

IV. Testing for Equality of Winning Probabilities

We first test, for each of the 40 point-game "experiments" in our data set, the hypothesis that the server's winning probabilities were the same for left and right serves. We represent each experiment's data as having been generated by random draws from two binomial processes—a left process, which determines the winner of the point if the server has served to the left; and a right process, which determines who wins the point if the serve was to the right. The processes' binomial parameters are not known, and they might differ across the 40 experiments. We first consider each experiment in isolation: in each one, our null hypothesis is that the left and right processes' binomial parameters are the same—i.e., that the server's winning probabilities in that point game were the same for left serves as for right serves.

We use Karl Pearson's chi-square goodness-of-fit test of equality of two distributions (see, for example, p. 449 of Alexander M. Mood et al., 1974). We index the 40 point-game experiments by i ($i = 1, \ldots, 40$). For each experiment i, our *null hypothesis* is that $p_L^i = p_R^i$, or equivalently, that there is a number p^i such that $p_L^i = p^i$ and $p_R^i = p^i$. If the null hypothesis is true, then the Pearson test statistic is distributed asymptotically as chi-square with two

[8] There is some possibility that selecting only long (and thus close) matches could introduce a sample selection bias: matches might be long partly *because* both players are playing as the equilibrium predicts.

[9] We are interested in the relation between (first) serve direction and whether the server ultimately wins the point. Therefore, for example, each of the following cases would yield an increment in both the number of serves to L and the number of points won when the serve is to L: (a) when a first serve is to L and the serve is good and the server wins the point; and (b) when a first serve is to L and the serve is a fault and the server wins the point following the second serve, which could be in any direction.

TABLE 1—TESTING FOR EQUALITY OF WINNING PROBABILITIES IN TENNIS DATA

			Serves			Mixture		Points Won		Win Rates		Pearson	
Match	Server	Court	L	R	Total	L	R	L	R	L	R	statistic	p-value
74Wimbldn	**Rosewall**	Ad	37	37	74	0.50	0.50	25	26	0.68	0.70	0.063	0.802
74Wimbldn	**Rosewall**	Deuce	70	5	75	0.93	0.07	50	3	0.71	0.60	0.294	0.588
74Wimbldn	Smith	Ad	66	10	76	0.87	0.13	45	7	0.68	0.70	0.013	0.908
74Wimbldn	Smith	Deuce	53	29	82	0.65	0.35	33	14	0.62	0.48	1.499	0.221
80Wimbldn	**Borg**	Ad	19	73	92	0.21	0.79	11	50	0.58	0.68	0.758	0.384
80Wimbldn	**Borg**	Deuce	37	62	99	0.37	0.63	26	41	0.70	0.66	0.182	0.670
80Wimbldn	McEnroe	Ad	45	40	85	0.53	0.47	27	26	0.60	0.65	0.226	0.635
80Wimbldn	McEnroe	Deuce	44	44	88	0.50	0.50	28	32	0.64	0.73	0.838	0.360
80USOpen	**McEnroe**	Ad	39	40	79	0.49	0.51	23	30	0.59	0.75	2.297	0.130
80USOpen	**McEnroe**	Deuce	51	32	83	0.61	0.39	31	18	0.61	0.56	0.167	0.683
80USOpen	Borg	Ad	29	47	76	0.38	0.62	17	30	0.59	0.64	0.206	0.650
80USOpen	Borg	Deuce	30	50	80	0.38	0.63	20	26	0.67	0.52	1.650	0.199
82Wimbldn	**Connors**	Ad	32	46	78	0.41	0.59	16	32	0.50	0.70	3.052	0.081**
82Wimbldn	**Connors**	Deuce	76	15	91	0.84	0.16	51	8	0.67	0.53	1.042	0.307
82Wimbldn	McEnroe	Ad	32	39	71	0.45	0.55	23	24	0.72	0.62	0.839	0.360
82Wimbldn	McEnroe	Deuce	35	44	79	0.44	0.56	24	30	0.69	0.68	0.001	0.970
84French	**Lendl**	Ad	33	34	67	0.49	0.51	18	21	0.55	0.62	0.359	0.549
84French	**Lendl**	Deuce	26	45	71	0.37	0.63	19	31	0.73	0.69	0.139	0.710
84French	McEnroe	Ad	38	29	67	0.57	0.43	23	18	0.61	0.62	0.016	0.898
84French	McEnroe	Deuce	42	30	72	0.58	0.42	21	20	0.50	0.67	1.983	0.159
87Australn	**Edberg**	Ad	47	22	69	0.68	0.32	29	12	0.62	0.55	0.318	0.573
87Australn	**Edberg**	Deuce	19	56	75	0.25	0.75	12	40	0.63	0.71	0.456	0.499
87Australn	Cash	Ad	38	27	65	0.58	0.42	19	14	0.50	0.52	0.022	0.883
87Australn	Cash	Deuce	39	29	68	0.57	0.43	25	16	0.64	0.55	0.554	0.457
88Australn	**Wilander**	Ad	32	36	68	0.47	0.53	20	25	0.63	0.69	0.365	0.546
88Australn	**Wilander**	Deuce	20	56	76	0.26	0.74	16	35	0.80	0.63	2.045	0.153
88Australn	Cash	Ad	40	23	63	0.63	0.37	22	13	0.55	0.57	0.014	0.907
88Australn	Cash	Deuce	37	37	74	0.50	0.50	19	25	0.51	0.68	2.018	0.155
88Masters	**Becker**	Ad	50	26	76	0.66	0.34	30	18	0.60	0.69	0.626	0.429
88Masters	**Becker**	Deuce	53	31	84	0.63	0.37	38	20	0.72	0.65	0.472	0.492
88Masters	Lendl	Ad	55	21	76	0.72	0.28	43	15	0.78	0.71	0.383	0.536
88Masters	Lendl	Deuce	46	38	84	0.55	0.45	24	23	0.52	0.61	0.589	0.443
95USOpen	**Sampras**	Ad	20	37	57	0.35	0.65	12	28	0.60	0.76	1.524	0.217
95USOpen	**Sampras**	Deuce	33	26	59	0.56	0.44	20	22	0.61	0.85	4.087	0.043*
95USOpen	Agassi	Ad	39	16	55	0.71	0.29	29	13	0.74	0.81	0.298	0.585
95USOpen	Agassi	Deuce	30	29	59	0.51	0.49	17	17	0.57	0.59	0.023	0.879
97USOpen	**Korda**	Ad	55	19	74	0.74	0.26	42	16	0.76	0.84	0.513	0.474
97USOpen	**Korda**	Deuce	52	30	82	0.63	0.37	38	19	0.73	0.63	0.852	0.356
97USOpen	Sampras	Ad	33	51	84	0.39	0.61	21	32	0.64	0.63	0.007	0.934
97USOpen	Sampras	Deuce	50	43	93	0.54	0.46	33	28	0.66	0.65	0.008	0.929
	Totals		1,622	1,404	3,026	0.54	0.46	1,040	918	0.64	0.65	30.801	0.852

* Indicates rejection at the 5-percent level of significance.
** Indicates rejection at the 10-percent level of significance.

degrees of freedom if p^i is known, or with one degree of freedom if p^i must be estimated from the data, as in our case.

Table 1 reports the results of the Pearson test. For each of the 40 point-game experiments, the two columns labeled "Pearson sta-

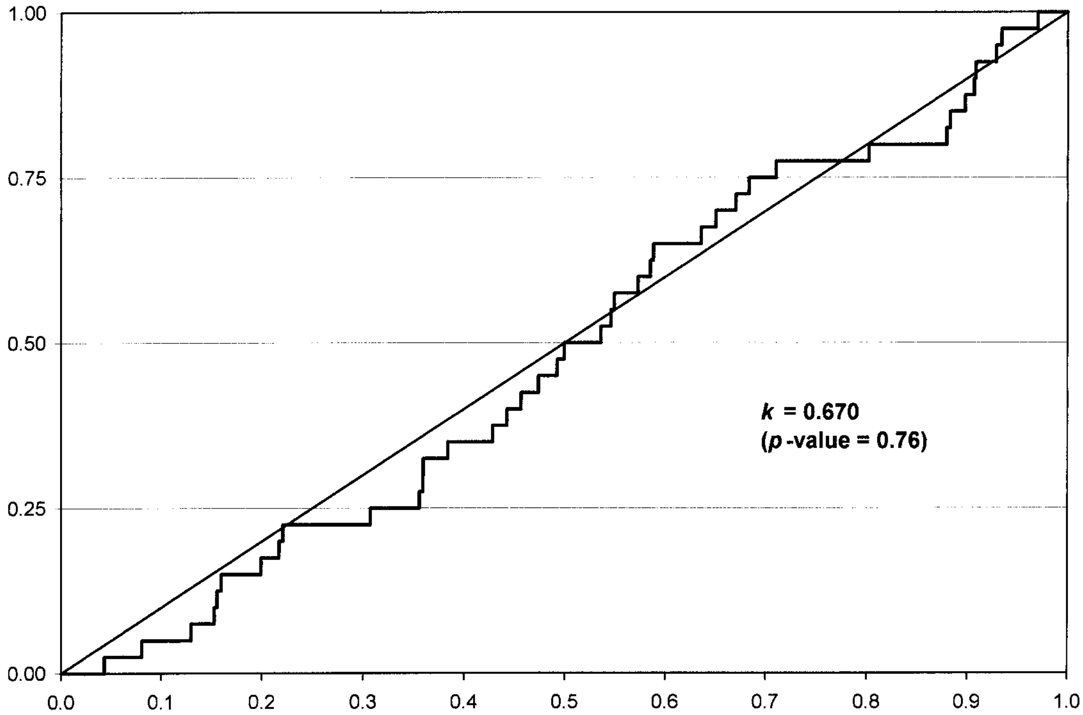


FIGURE 2. WIN RATES IN TENNIS: KOLMOGOROV TEST

tistic" and "p-value," at the right-hand side of the table, report the value of the test statistic, Q^i, along with its associated p-value. In only one of our 40 point-game experiments (Sampras serving to Agassi in the deuce court in 1995) do we find the null hypothesis rejected at the 5-percent level (a p-value less that 0.05), and for only one other point game (Connors serving to McEnroe in the ad court in 1982) do we reject at the 10-percent level. Note that with 40 point games, the expected number of individual rejections according to the theory (i.e., when the null hypothesis is true) is two rejections at the 5-percent level and four at the 10-percent level. Considering simply the number of 5-percent and 10-percent rejections, then, the tennis data appear quite consistent with the theory.

This suggests a test of the *joint* hypothesis that the data from *all 40* experiments were generated by equilibrium play. We apply Pearson's test to the joint hypothesis that $p_L^i = p_R^i$ for *each one* of the experiments $i = 1, \ldots, 40$ (but allowing the parameters p_L^i and p_R^i to vary across experiments i). The test statistic for the Pearson joint test is simply the sum of the test statistics Q^i in the 40 individual tests we have just described, which under the null hypothesis is distributed as chi-square with 40 degrees of freedom. For our tennis data, the value of this test statistic is 30.801 and the associated p-value is 0.852. Clearly, we cannot reject this joint hypothesis at any reasonable level of significance.

We have observed, above, that in the 40 individual tests, the data yield slightly *fewer* rejections of the null hypothesis than one would expect to obtain when the theory is correct—i.e., when the joint null hypothesis is true. We develop this idea further, to obtain a more informative assessment of the data's conformity with the theory. We consider all 40 point-game experiments, and we compare the observed distribution of the 40 Q^i values with the distribution predicted by the theory. Recall that under the joint null hypothesis ($p_L^i = p_R^i$ for each experiment i) the Pearson

TABLE 2—TESTING FOR EQUALITY OF WINNING PROBABILITIES IN O'NEILL'S DATA

		Mixtures		Win Rates			
Pair	Player	Joker	Non-Joker	Joker	Non-Joker	Pearson Q	p-value
1	1	0.181	0.819	0.211	0.430	3.156	0.076**
	2	0.352	0.648	0.892	0.456	19.139	0.000*
2	1	0.438	0.562	0.391	0.220	3.631	0.057**
	2	0.552	0.448	0.690	0.723	0.142	0.706
3	1	0.543	0.457	0.526	0.229	9.667	0.002*
	2	0.552	0.448	0.483	0.766	8.749	0.003*
4	1	0.333	0.667	0.829	0.214	36.167	0.000*
	2	0.724	0.276	0.618	0.483	1.587	0.208
5	1	0.467	0.533	0.388	0.304	0.822	0.365
	2	0.448	0.552	0.596	0.707	1.424	0.233
6	1	0.390	0.610	0.463	0.391	0.544	0.461
	2	0.448	0.552	0.596	0.569	0.076	0.782
7	1	0.305	0.695	0.531	0.452	0.559	0.454
	2	0.352	0.648	0.541	0.515	0.064	0.800
8	1	0.324	0.676	0.412	0.493	0.609	0.435
	2	0.295	0.705	0.548	0.527	0.040	0.841
9	1	0.295	0.705	0.290	0.392	0.976	0.323
	2	0.343	0.657	0.750	0.580	2.971	0.085**
10	1	0.419	0.581	0.364	0.410	0.229	0.632
	2	0.410	0.590	0.628	0.597	0.103	0.748
11	1	0.305	0.695	0.313	0.425	1.176	0.278
	2	0.371	0.629	0.744	0.530	4.686	0.030*
12	1	0.486	0.514	0.490	0.593	1.108	0.292
	2	0.429	0.571	0.444	0.467	0.051	0.821
13	1	0.267	0.733	0.536	0.364	2.514	0.113
	2	0.533	0.467	0.732	0.429	9.959	0.002*
14	1	0.305	0.695	0.344	0.521	2.794	0.095**
	2	0.229	0.771	0.542	0.531	0.009	0.926
15	1	0.457	0.543	0.313	0.333	0.052	0.820
	2	0.371	0.629	0.615	0.712	1.048	0.306
16	1	0.438	0.562	0.304	0.373	0.539	0.463
	2	0.381	0.619	0.650	0.662	0.015	0.904
17	1	0.362	0.638	0.368	0.358	0.011	0.917
	2	0.410	0.590	0.674	0.613	0.416	0.519
18	1	0.390	0.610	0.488	0.484	0.001	0.973
	2	0.410	0.590	0.535	0.500	0.124	0.725
19	1	0.324	0.676	0.500	0.338	2.534	0.111
	2	0.505	0.495	0.679	0.538	2.186	0.139
20	1	0.429	0.571	0.600	0.317	8.386	0.004*
	2	0.495	0.505	0.481	0.642	2.755	0.097**
21	1	0.371	0.629	0.436	0.500	0.404	0.525
	2	0.324	0.676	0.500	0.535	0.114	0.735

TABLE 2—*Continued.*

		Mixtures		Win Rates			
Pair	Player	Joker	Non-Joker	Joker	Non-Joker	Pearson Q	p-value
22	1	0.457	0.543	0.354	0.439	0.774	0.379
	2	0.343	0.657	0.528	0.638	1.191	0.275
23	1	0.162	0.838	0.471	0.443	0.043	0.835
	2	0.419	0.581	0.818	0.361	21.641	0.000*
24	1	0.257	0.743	0.519	0.487	0.079	0.779
	2	0.371	0.629	0.641	0.424	4.609	0.032*
25	1	0.333	0.667	0.486	0.257	5.486	0.019*
	2	0.590	0.410	0.726	0.581	2.383	0.123
						167.741	0.000*

* 10 rejections at 5 percent.
** 15 rejections at 10 percent.

statistic Q^i is asymptotically distributed as chi-square-1 for each i. In other words, each experiment yields an independent draw, Q^i, from the chi-square-1 distribution, and thus (under the joint null hypothesis) the 40 Q^i values in Table 1 should be 40 such chi-square draws. Equivalently, the p-values associated with the realized Q^i values (also in Table 1) should have been 40 draws from the uniform distribution $U[0, 1]$.

A simple visual comparison of the observed distribution with the theoretically predicted distribution is provided in Figure 2, in which the empirical cumulative distribution function (the c.d.f.) of the 40 observed p-values is juxtaposed with the theoretical c.d.f., the distribution the p-values would have been drawn from if the data were generated according to the theory—i.e., the uniform distribution, for which the c.d.f. is the 45° line in Figure 2. The empirical and the theoretically predicted distributions depicted in Figure 2 are strikingly close to one another.

We formalize this comparison of the two distributions via the Kolmogorov-Smirnov (KS) test, which allows one to test the hypothesis that an empirical distribution of observed values was generated by draws from a specified ("hypothesized") distribution. In addition to its appealing visual interpretation, as in Figure 2, the KS test is also more powerful than the Pearson joint test against many alternative hypotheses about how the data were generated.[10]

In performing the KS test for the tennis data, the hypothesized c.d.f. for the p-values is the uniform distribution, $F(x) = x$ for $x \epsilon [0, 1]$. Denoting the empirical distribution of the 40 p-values in Table 1 by $\hat{F}(x)$, the KS test statistic is $K = \sqrt{40}\ \sup_{x \in [0,1]} |\hat{F}(x) - x|$, which has a known distribution (see p. 509 of Mood et al., 1974). For the tennis data in Table 1, we have $K = 0.670$, with a p-value of 0.76, far toward the opposite end of the distribution from the rejection region. This data, in other words, is *typical* of the data that minimax play would produce: minimax play would generate a value of K at least this large 76 percent of the time, and a "better" (smaller) value of K only 24 percent of the time.

A. *Applying Our Tests to Experimental Data*

In experiments on mixed-strategy play, observed play adhered most closely to the equilibrium prediction in Barry O'Neill's (1987) experiment.[11] When we apply the same statistical tests to O'Neill's data as we

[10] See Walker and Wooders (1999 fn. 19) for a simple illustration of this.

[11] O'Neill's ingenious experimental design avoided several weaknesses he had identified in prior tests of the theory.

have to our tennis data, the difference is striking.

In O'Neill's experiment 25 pairs of subjects repeatedly played a simple two-person game in which each player chooses one of four cards: Ace, Two, Three, or Joker. The game has a unique Nash equilibrium: each player chooses the Joker with probability 0.4 and chooses each number card with probability 0.2. O'Neill's subjects all played the game 105 times, each subject always playing against the same opponent.[12] O'Neill awarded the winner of each play a nickel and the loser nothing (a total of $5.25 per pair of subjects).

In order to compare our binary-choice data with O'Neill's data, we pool the three number cards, which are strategically equivalent, into a single action, "non-Joker," and focus our analysis on the binary choice of a Joker or a non-Joker card. We index the subjects by $i \in \{1, \ldots, 50\}$. Of course each subject's choices were observable in O'Neill's experiment, so we can use each of the 50 subjects' win rates to test whether his winning probabilities were the same for his plays of the Joker and the non-Joker cards. We conduct the same tests as we have already carried out above for the tennis data. Table 2 contains the observed mixtures and win rates, and the corresponding values of the test statistic and its p-values.

In the 50 individual tests (in each of which the null hypothesis is that the subject's Joker and non-Joker winning probabilities are the same), we obtain 10 rejections at the 5-percent level and 15 rejections at the 10-percent level.[13] In order to test the *joint* hypothesis that the winning probability is the same for Joker and non-Joker cards for *every* subject (but possibly different across subjects), we simply sum the 50 values of the test statistic to obtain the statistic $\sum_{i=1}^{50} Q^i$, just as we described above for the tennis data. This statistic is asymptotically distributed chi-square with 50 degrees of freedom under the null hypothesis. The value of the statistic is 167.741 and the associated p-value is 1.239×10^{-14}; hence the joint null hypothesis is rejected at virtually any level of significance, in sharp contrast to the large p-value (0.852) obtained in the parallel test on the tennis data.

Figure 3 is the analogue for O'Neill's data of Figure 2 for the tennis data: it depicts the empirical distribution and the hypothesized (i.e., uniform) distribution of the p-values for the tests of equality of winning probabilities in O'Neill's data. The value of the KS test statistic is $K = 1.704$, with a p-value of 0.006. Hence the KS test rejects the null hypothesis at significance levels as low as 1 percent, again a sharp contrast with the tennis data, in which the corresponding p-value is 0.76. Comparison of Figures 2 and 3 provides a striking visual picture of the difference between the two data sets' conformity with the theory of mixed-strategy equilibrium. The p-values in the tennis data are distributed almost exactly uniformly, as the theory predicts they should be, but the p-values are far from uniform for O'Neill's data.

B. *The Power Of Our Tests*

In order to evaluate the power of our tests, we concentrate our attention on the Pearson joint test for equality of the server's left and right winning probabilities. Using the numerical example in Figure 1,[14] we formulate a parametric class of plausible alternative hypotheses and we conduct Monte Carlo simulations to evaluate the power of the Pearson

[12] Regardless of the players' risk attitudes, the unique equilibrium of the repeated O'Neill game consists of the players mixing independently at each stage according to the stage game's equilibrium mixture. This follows from results in Wooders and Jason Shachat (2001), who study sequential play of stage games in which each stage game has only two possible outcomes.

[13] James N. Brown and Robert W. Rosenthal (1990) have provided an extensive statistical analysis of O'Neill's data, including direct tests using subjects' empirical mixtures, and they obtain similar levels of rejection.

[14] Of course, as we have already pointed out, the actual probability payoffs in the point games are not observable, and they surely differ from one "experiment" to another. However, the point-game example in Figure 1 captures some of the key aggregate features of the actual data in our tennis matches: in the game's equilibrium the server serves to the receiver's left with mixture probability 0.53⅓, while in the data 53.5 percent of all first serves are to the left; and the game's value (i.e., the probability that the server will win the point) is 0.65, while in the data the servers won 64.7 percent of all points.

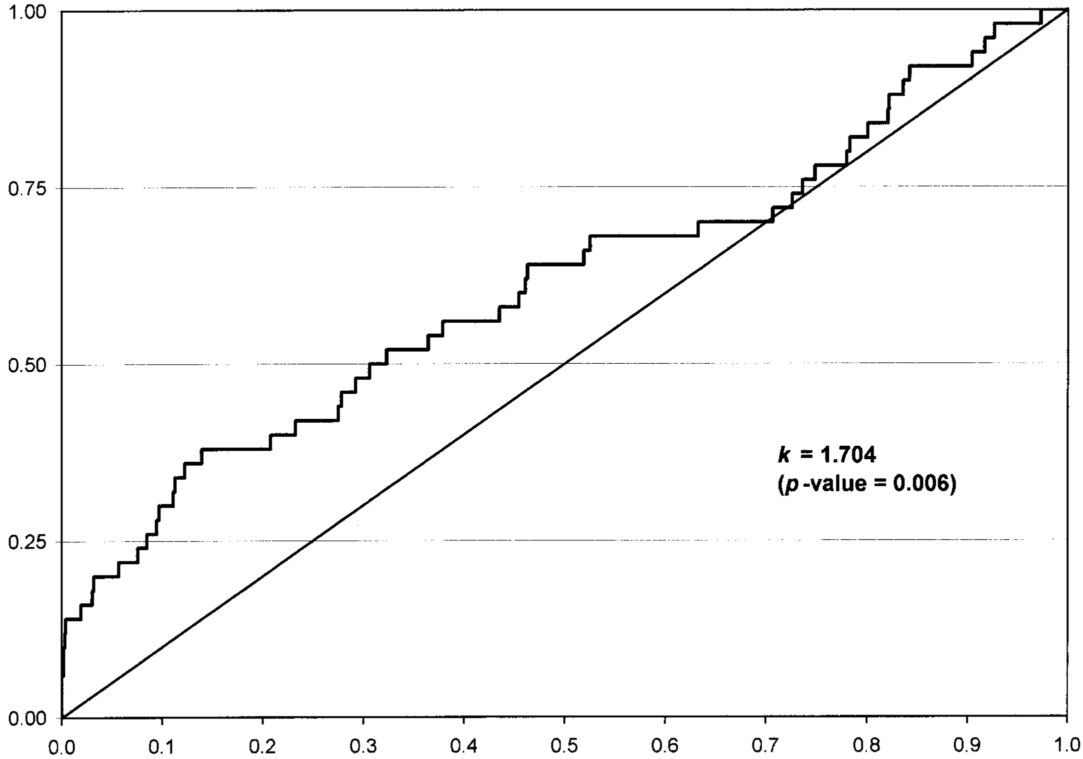


FIGURE 3. WIN RATES IN O'NEILL: KOLMOGOROV TEST

test to reject the null hypothesis when any of these alternative hypotheses is true.

Our null hypothesis in each experiment (viz., that the server's left and right winning probabilities are the same) is a consequence of our assumption that the receiver is playing his minimax mixture. But the receiver may actually be mixing his choices in some other proportions. Let θ denote the proportion of the points on which the receiver chooses left; the null hypothesis for a given experiment is thus that $\theta = \frac{2}{3}$, and alternative values of θ comprise the alternative hypotheses we will consider. For any value of θ, the server's winning probabilities p_L and p_R are given by

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

and

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

Under the joint null hypothesis that in a data set with 40 experiments each receiver follows his minimax mixture, the Pearson test statistic $\Sigma_{i=1}^{40} Q^i$ is asymptotically distributed as chi-square with 40 degrees of freedom.

At the 5-percent significance level, the Pearson joint test consists of rejecting the null hypothesis if $\Sigma_{i=1}^{40} Q^i$ exceeds the critical value 55.75. The power of this test against an alternative value of θ is defined as the probability of rejecting the joint null hypothesis when the alternative value of θ is the true value. But for values of θ different from $\theta_0 = \frac{2}{3}$, we have $p_L \neq p_R$, and hence the distribution of the Pearson test statistic $\Sigma_{i=1}^{40} Q^i$ is not known. We have used Monte Carlo methods to estimate the power of the test against alternative values of θ.[15] The power function is depicted in

[15] For a given, fixed value of θ, data was randomly generated for 40 experiments; the test statistic was computed and compared to the critical value 55.75, and the null hypothesis was thus either rejected or it was not. This process was repeated 100,000 times, with the empirical frequency of rejection then used as the estimate of the test's

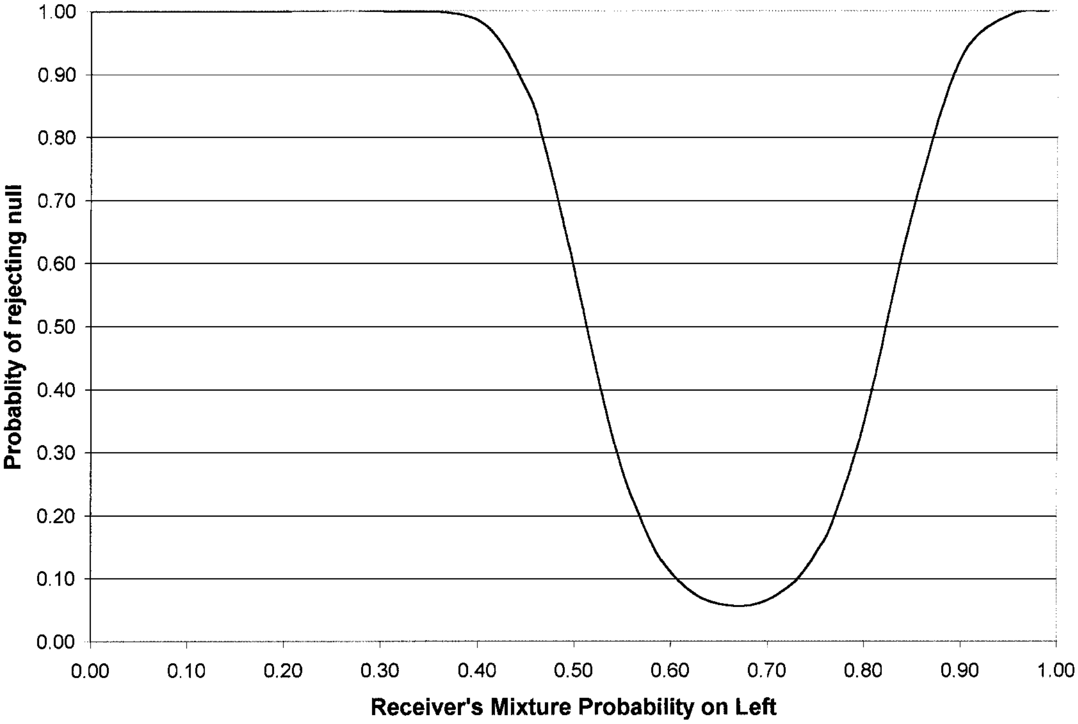


FIGURE 4. THE POWER FUNCTION

Figure 4. One can see that the Pearson joint test has significant ability to reject the null hypothesis when the true value of θ differs significantly from $\theta_0 = \frac{2}{3}$. For example, if $\theta = 0.5$—i.e., if the receivers actually choose left and right with equal probability—the joint null hypothesis is rejected with probability 0.58. Thus, the test has fairly high power, even though the server's winning probabilities are not very different (they are $p_L = 0.685$ and $p_R = 0.610$). Under the alternative hypothesis that $\theta = 0.4$, i.e., that the receiver chooses left with probability 0.4, the power of the test rises dramatically to 0.98.

V. Serial Independence

First we test individually, for each of the 40 "experiments" in our data set, the hypothesis that the server's choices were serially independent. Let $s^i = (s^i_1, \ldots, s^i_{n^i_L + n^i_R})$ be the list of first-serve directions in experiment i, in the order in which they occurred, where $s^i_n \in \{L, R\}$ is the direction of the n^{th} first serve, and where n^i_L and n^i_R are the number of first serves to the left and to the right. Our test of serial independence is based on the number of runs in the list s^i, which we denote by r^i. (A *run* is a maximal string of consecutive identical symbols, either all L's or all R's.[16]) We reject the hypothesis of serial independence if there are either "too many" runs or "too few" runs. Too many runs suggests negative correlation in the choice of direction: the runs tend to be too short, and thus the server is changing direction too often for his choices to have been randomly generated. Too few runs suggests that the server's choices are positively correlated: the server is not changing direction often enough to be consistent with

power under θ, i.e., the probability of rejecting when θ is true. This Monte Carlo estimation of the test's power was performed for many values of θ.

[16] For example, the sequence $s = (\text{L, L, R, L})$ has three runs. We omit serves to the center.

randomness, resulting in runs that tend to be too long.

Under the null hypothesis of serial independence, the probability that there are exactly r runs in a list made up of n_L and n_R occurrences of L and R is known (see, for example, Jean Dickinson Gibbons and Subhabrata Chakraborti, 1992). Denote this probability by $f(r;\ n_L,\ n_R)$, and let $F(r;\ n_L,\ n_R)$ denote the value of the associated c.d.f., i.e., $F(r;\ n_L,\ n_R) = \Sigma_{k=1}^{r} f(k;\ n_L,\ n_R)$, the probability of obtaining r or fewer runs. At the 5-percent significance level, the null hypothesis of serial independence in experiment i is rejected if either $F(r^i;\ n_L^i,\ n_R^i) < 0.025$ or $1 - F(r^i - 1;\ n_L^i,\ n_R^i) < 0.025$, i.e., if the probability of r^i or fewer runs is less than 0.025 or the probability of r^i or more runs is less than 0.025.

Table 3 shows the data and the results for our tests of serial independence. For each of the 40 point-game experiments, the columns L, R, and Total give the number of first serves in each direction and the total number of first serves. The Runs column indicates the number of runs, r^i, in the list s^i of first serves (the lists are not shown). The columns $F(r - 1)$ and $F(r)$ give the value of the c.d.f. in experiment i for $r^i - 1$ and r^i runs, respectively. At the 5-percent significance level, the null hypothesis is rejected in five of the 40 experiments (the expected number of 5-percent rejections is only two). In three cases the null hypothesis is rejected because there are too many runs, and in two cases the rejection is because there are too few runs.

To test the joint hypothesis that first serves are serially independent in *each* of the 40 experiments, we again employ the Kolmogorov-Smirnov goodness-of-fit test. The KS test requires that the sequence of random variables of interest be independently and identically distributed, with a *continuous* cumulative distribution function. Hence, the KS test cannot be applied directly to the values in either column $F(r - 1)$ or column $F(r)$, since these values are neither identically distributed (the distribution of r^i depends on n_L^i and n_R^i) nor continuously distributed. We circumvent these difficulties by constructing, for each experiment i, the (random) statistic t^i given by a draw from the uniform distribution $U[F(r^i - 1;\ n_L^i,\ n_R^i),\ F(r^i;\ n_L^i,\ n_R^i)]$. A particular realization of this statistic for each experiment is given in the right-most column of Table 3. Under the null hypothesis of serial independence in experiment i, the statistic t^i is distributed $U[0,\ 1]$.[17]

The empirical c.d.f. of the realized values $t^1, \ldots, t^{40}$ in Table 3 is depicted in Figure 5. The value of the KS test statistic is $K = 1.948$,[18] with a p-value of 0.001. Hence, we reject the null hypothesis that in all 40 experiments the first serves were serially independent. Figure 5 and Table 3 show that there tend to be too many large values of t^i, i.e., too many runs, relative to the null hypothesis.

The finding that even the best tennis players typically switch from one action to another too often is perhaps not surprising. There is overwhelming experimental evidence that when people try to generate "random" sequences they generally "switch too often" to be consistent with randomly generated choices (W. A. Wagenaar, 1972).

A. *Serial Independence in O'Neill's Data*

Table 4 shows the data and the results of tests for serial independence in O'Neill's experiment. We distinguish only between Jokers and non-Jokers when counting runs. For each of O'Neill's 50 subjects, the columns J and N in Table 4 indicate the number of times the subject chose Joker and non-Joker (out of 105 plays altogether), and the Runs column indicates the number of runs in the subject's list of choices. At the 5-percent significance level, the null hypothesis that play is serially independent is rejected for 15 subjects (the expected number is only 2.5). In 13 of the 15 rejections there are too many runs, and in the other two there are too few runs.

The values in the right-most column of Table 4 are, for each subject i, a realized value of the test statistic t^i constructed as described above. The empirical cumulative distribution of these t^i values is shown in Figure 6. The value of the KS test statistic is $K = 2.503$, with a p-value of 0.000007. Hence we reject the joint null hypoth-

[17] A proof is contained in footnote 24 of Walker and Wooders (1999).

[18] This "randomized" test was performed many times. While there was of course variation in the 40 t^i values across trials, there was only slight variation in the value of K and in the p-value, at the third decimal place and beyond.

TABLE 3—RUNS TESTS ON TENNIS DATA

			Serves			Runs			
Match	Server	Court	L	R	Total	r^i	$F(r^i - 1)$	$F(r^i)$	$U[F(r^i - 1), F(r^i)]$
74Wimbldn	**Rosewall**	Ad	37	37	74	43	0.854	0.901	0.866
74Wimbldn	**Rosewall**	Deuce	70	5	75	11	0.349	1.000	0.804
74Wimbldn	Smith	Ad	66	10	76	21	0.812	1.000	0.823
74Wimbldn	Smith	Deuce	53	29	82	43	0.832	0.892	0.852
80Wimbldn	**Borg**	Ad	19	73	92	33	0.633	0.788	0.757
80Wimbldn	**Borg**	Deuce	37	62	99	52	0.817	0.866	0.855
80Wimbldn	McEnroe	Ad	45	40	85	44	0.512	0.599	0.553
80Wimbldn	McEnroe	Deuce	44	44	88	49	0.774	0.832	0.818
80USOpen	**McEnroe**	Ad	39	40	79	38	0.249	0.326	0.298
80USOpen	**McEnroe**	Deuce	51	32	83	36	0.131	0.185	0.142
80USOpen	Borg	Ad	29	47	76	42	0.873	0.916	0.912
80USOpen	Borg	Deuce	30	50	80	43	0.829	0.887	0.844
82Wimbldn	**Connors**	Ad	32	46	78	49	0.990*	0.995	0.994
82Wimbldn	**Connors**	Deuce	76	15	91	31	0.958**	1.000	0.999
82Wimbldn	McEnroe	Ad	32	39	71	36	0.437	0.533	0.520
82Wimbldn	McEnroe	Deuce	35	44	79	36	0.152	0.212	0.183
84French	**Lendl**	Ad	33	34	67	41	0.931	0.958	0.938
84French	**Lendl**	Deuce	26	45	71	41	0.955**	0.976	0.963
84French	McEnroe	Ad	38	29	67	40	0.921	0.952	0.947
84French	McEnroe	Deuce	42	30	72	45	0.982*	0.991	0.984
87Australn	**Edberg**	Ad	47	22	69	40	0.994*	0.997	0.997
87Australn	**Edberg**	Deuce	19	56	75	29	0.374	0.519	0.505
87Australn	Cash	Ad	38	27	65	40	0.964**	0.980	0.968
87Australn	Cash	Deuce	39	29	68	37	0.711	0.791	0.725
88Australn	**Wilander**	Ad	32	36	68	38	0.739	0.813	0.795
88Australn	**Wilander**	Deuce	20	56	76	29	0.265	0.389	0.275
88Australn	Cash	Ad	40	23	63	29	0.316	0.424	0.364
88Australn	Cash	Deuce	37	37	74	28	0.007	0.013*	0.010
88Masters	**Becker**	Ad	50	26	76	38	0.724	0.796	0.783
88Masters	**Becker**	Deuce	53	31	84	45	0.847	0.900	0.890
88Masters	Lendl	Ad	55	21	76	32	0.515	0.607	0.539
88Masters	Lendl	Deuce	46	38	84	43	0.489	0.577	0.506
95USOpen	**Sampras**	Ad	20	37	57	25	0.231	0.335	0.245
95USOpen	**Sampras**	Deuce	33	26	59	22	0.011	0.021*	0.019
95USOpen	Agassi	Ad	39	16	55	29	0.943	0.980	0.968
95USOpen	Agassi	Deuce	30	29	59	24	0.032	0.058	0.052
97USOpen	**Korda**	Ad	55	19	74	28	0.301	0.389	0.323
97USOpen	**Korda**	Deuce	52	30	82	43	0.793	0.859	0.842
97USOpen	Sampras	Ad	33	51	84	35	0.065	0.101	0.079
97USOpen	Sampras	Deuce	50	43	93	41	0.079	0.114	0.087

* Indicates rejection at the 5-percent level.

** Indicates rejection at the 10-percent level.

esis that each of the 50 subjects' choices were serially independent in O'Neill's experiment. Just as in the tennis data, there are generally too many runs for the joint null hypothesis to be true. Comparing the empirical cumulative distribution functions in Figures 5 and 6 suggests that while play is negatively correlated in both the tennis data and O'Neill's experimental data

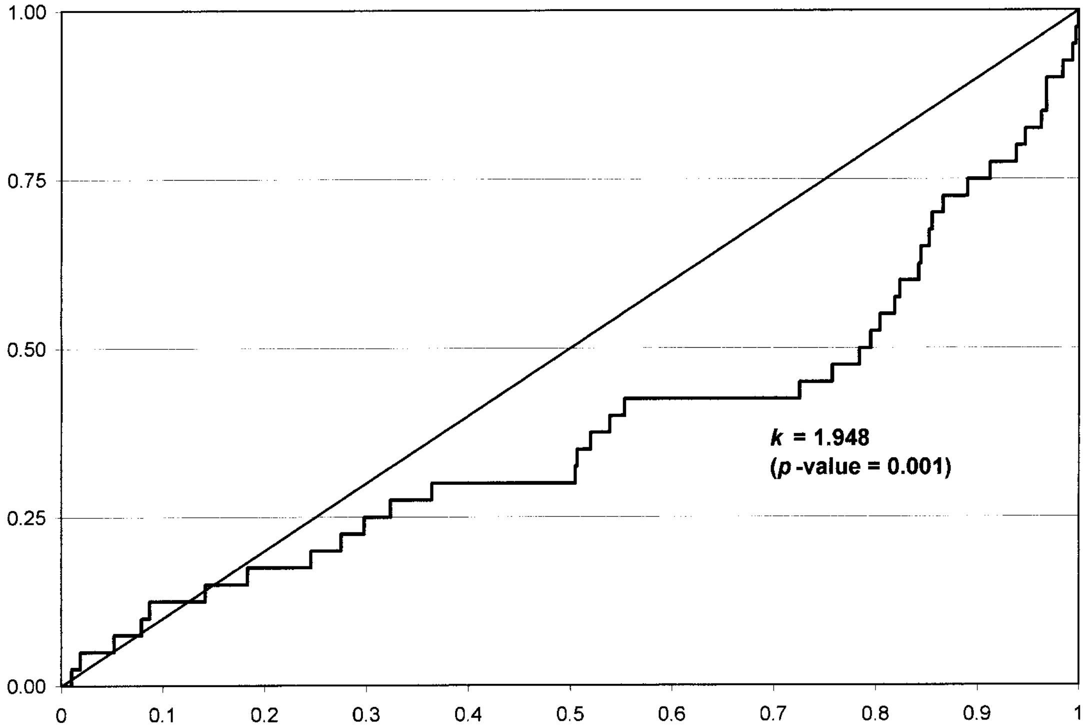


FIGURE 5. RUNS IN TENNIS DATA: KOLMOGOROV TEST

(generally too much switching between choices in both cases), the correlation is clearly less in the tennis data.

Thus, just as with our tests using players' win rates, the tests for randomness (and serial correlation in particular) reveal a striking difference between the theory's consistency with the data for top tennis players and its consistency with the data from experiments.

VI. Concluding Remarks

The theory of mixed-strategy equilibrium has not been consistent with the empirical evidence gathered through more than 40 years of experiments involving human subjects. Conversely, the theory has performed far better in explaining the play of top professional tennis players in our data set. We do not view these results as an indictment of the many experiments that have been conducted to test for equilibrium play: the experiments have established convincingly that when unpredictable play is called for, inexperienced players will not generally mix in the equilibrium proportions. Nor do we mean to suggest that the theory applies only to people who have developed years of experience in a particular strategic situation. There is a spectrum of experience and expertise, with novices (such as typical experimental subjects) at one extreme and world-class tennis players at the other. The theory applies well (but not perfectly) at the "expert" end of the spectrum, in spite of its failure at the "novice" end. There is a very large gulf between the two extremes, and little, if anything, is presently known about how to place a given strategic situation along this spectrum or about how to divide the spectrum into the portions on which current theory applies and the portions where a more general, or even a new, theory must be developed. The last ten years or

TABLE 4—RUNS TESTS ON O'NEILL'S DATA

		Choices		Runs			
Pair	Player	J	N	r^i	$F(r^i - 1)$	$F(r^i)$	$U[F(r^i-1), F(r^i)]$
1	1	19	86	34	0.688	0.753	0.718
	2	37	68	47	0.297	0.381	0.337
2	1	46	59	66	0.995*	0.997	0.995
	2	58	47	51	0.315	0.389	0.323
3	1	57	48	57	0.748	0.807	0.750
	2	58	47	53	0.466	0.545	0.468
4	1	35	70	50	0.659	0.728	0.714
	2	76	29	55	0.999*	1.000	1.000
5	1	49	56	55	0.596	0.670	0.613
	2	47	58	62	0.956**	0.972	0.963
6	1	41	64	58	0.912	0.939	0.921
	2	47	58	34	0.000	0.000*	0.000
7	1	32	73	48	0.682	0.748	0.734
	2	37	68	68	1.000*	1.000	1.000
8	1	34	71	40	0.049	0.073	0.055
	2	31	74	54	0.985*	0.991	0.985
9	1	31	74	40	0.114	0.158	0.139
	2	36	69	63	0.999*	1.000	0.999
10	1	44	61	57	0.810	0.861	0.814
	2	43	62	57	0.830	0.878	0.866
11	1	32	73	40	0.086	0.122	0.090
	2	39	66	59	0.963**	0.978	0.973
12	1	51	54	58	0.786	0.839	0.831
	2	45	60	43	0.023	0.037**	0.027
13	1	28	77	38	0.131	0.179	0.173
	2	56	49	53	0.440	0.518	0.508
14	1	32	73	50	0.828	0.873	0.847
	2	24	81	46	0.990*	0.994	0.991
15	1	48	57	57	0.748	0.807	0.749
	2	39	66	59	0.963**	0.978	0.968
16	1	46	59	39	0.002	0.004*	0.003
	2	40	65	48	0.265	0.334	0.318
17	1	38	67	57	0.931	0.958	0.940
	2	43	62	68	0.999*	1.000	0.999
18	1	41	64	44	0.062	0.091	0.076
	2	43	62	45	0.070	0.102	0.080
19	1	34	71	56	0.975*	0.985	0.978
	2	53	52	58	0.784	0.837	0.836
20	1	45	60	70	1.000*	1.000	1.000
	2	52	53	79	1.000*	1.000	1.000
21	1	39	66	63	0.996*	0.998	0.998
	2	34	71	48	0.548	0.625	0.619

TABLE 4—*Continued.*

		Choices		Runs			
Pair	Player	J	N	r^i	$F(r^i - 1)$	$F(r^i)$	$U[F(r^i-1), F(r^i)]$
22	1	48	57	67	0.996*	0.998	0.998
	2	36	69	48	0.149	0.200	0.193
23	1	17	88	31	0.589	0.787	0.622
	2	44	61	65	0.994*	0.997	0.995
24	1	27	78	45	0.796	0.879	0.821
	2	39	66	58	0.944	0.963	0.963
25	1	35	70	52	0.804	0.854	0.824
	2	62	43	57	0.830	0.878	0.847

* Indicates rejection at the 5-percent level.
** Indicates rejection at the 10-percent level.

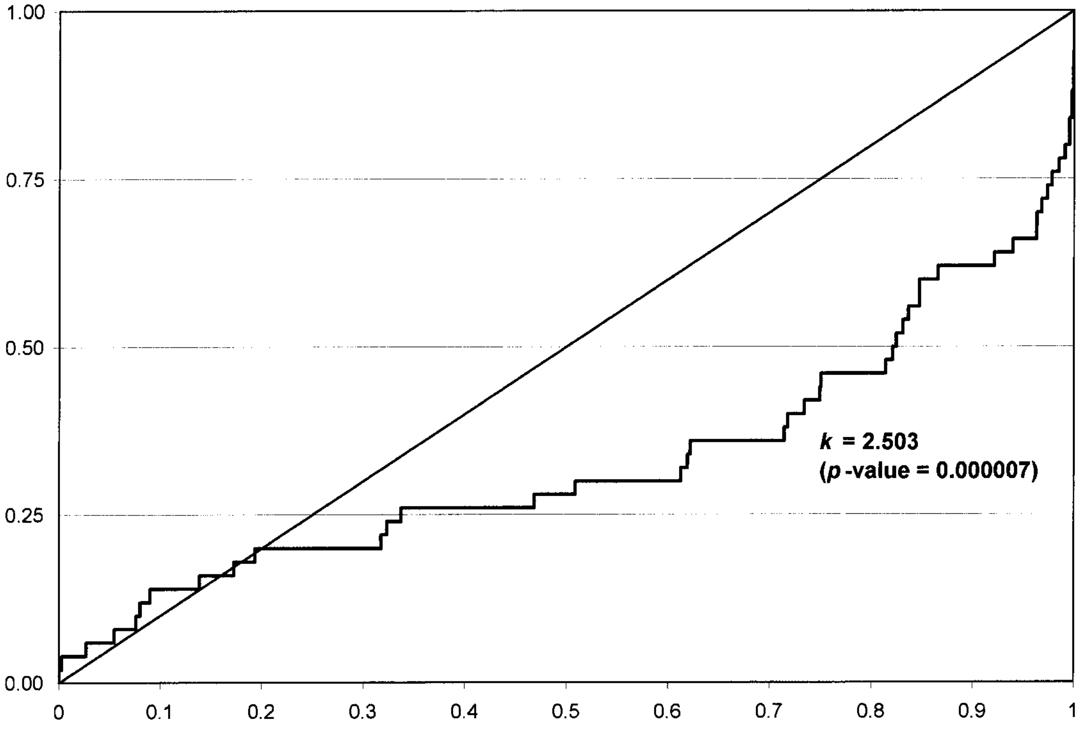


FIGURE 6. RUNS IN O'NEILL'S DATA: KOLMOGOROV TEST

so have seen the development of a large literature on out-of-equilibrium play, or "learning," in games, as well as alternative notions of equilibrium. This literature holds some promise for advancing our understanding of human behavior in strategic situations.

REFERENCES

Brown, James N. and Rosenthal, Robert W. "Testing the Minimax Hypothesis: A Reexamination of O'Neill's Game Experiment." *Econometrica*, September 1990, *58*(5), pp. 1065–81.

Dixit, Avinash and Nalebuff, Barry. *Thinking strategically: The competitive edge in business, politics, and everyday life.* New York: W.W. Norton, 1991.

Erev, Ido and Roth, Alvin E. "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria." *American Economic Review,* September 1998, *88*(4), pp. 848–81.

Feinstein, John. *Hard courts.* New York: Random House, 1991.

Gibbons, Jean Dickinson and Chakraborti, Subhabrata. *Nonparametric statistical inference.* New York: Marcel Dekker, 1992.

Mood, Alexander M.; Graybill, Franklin A. and Boes, Duane C. *Introduction to the theory of statistics.* New York: McGraw-Hill, 1974.

O'Neill, Barry. "Nonmetric Test of the Minimax Theory of Two-Person Zerosum Games." *Proceedings of the National Academy of Sciences*, April 1987, *84*, pp. 2106–09.

Wagenaar, W. A. "Generation of Random Sequences by Human Subjects: A Critical Survey of the Literature." *Psychological Bulletin*, 1972, *77*(2), pp. 65–72.

Walker, Mark and Wooders, John. "Minimax Play at Wimbledon." University of Arizona Working Paper No. 99-05, 1999.

______ . "Equilibrium Play in Matches: Binary Markov Games." University of Arizona Working Paper No. 00-12, 2000.

Wooders, John and Shachat, Jason. "On the Irrelevance of Risk Attitudes in Repeated Two Outcome Games." *Games and Economic Behavior*, February 2001, *34*(2), pp. 342–63.