Minimax Play at Wimbledon*

Mark Walker and John Wooders†

Department of Economics

University of Arizona

November 7, 1998

(Revised version: September 10, 1999)

Abstract

We use data from classic professional tennis matches to provide an empirical test of the theory of mixed strategy equilibrium. We find that the serve-and-return play of John McEnroe, Bjorn Borg, Boris Becker, Pete Sampras and others is consistent with equilibrium play. The same statistical tests soundly reject the assumption of equilibrium play in experimental data, including the data from Barry O'Neill's celebrated experiment.

*We are indebted to many people for helpful comments and discussions, but especially to Robert Aumann, Bruno Broseta, Bill Horrace, Tom Palfrey, Robert Rosenthal, Jason Shachat, and Vernon Smith, and to an anonymous referee for the *Review*. 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 #SBR-9710289 from the National Science Foundation.

†**mwalker@arizona.edu; jwooders@bpa.arizona.edu**

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 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 suggest an explanation for the theory's failure in experiments, and we use that explanation to suggest an alternative empirical test of the minimax hypothesis, a test based on field data from championship professional tennis matches. We then conduct such a test, and we find that the serve-and-return play of John McEnroe, Bjorn Borg, Boris Becker, Pete Sampras and other professional tennis players is largely consistent with the minimax hypothesis. When the same statistical tests are applied to data from some well known experiments on mixed strategy play, we soundly reject the assumption that the subjects were playing according to the mixed strategy equilibrium.

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

[1] See, for example, Figure 1 in Erev & Roth (1998) and their accompanying discussion, which describe twelve such experiments.

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're 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 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 fact, nearly all such examples in sports are empirical swamplands. In the classic confrontation between pitcher and batter, for example, there are many actions available (fastball, curve, slider, change-up, inside, outside, up in the strike zone, down in the zone, etc.), and the possible outcomes are even more numerous

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

[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."

(strike, ball, single, home run, fly ball, double-play grounder, etc). Difficulties clearly arise in modelling 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 modelling 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. The games always have a unique Nash equilibrium and the equilibrium requires each player to use a mixed strategy. (Because the games are two-person constant-sum games, the strategy a player uses in equilibrium is his unique minimax mixture.) However, the game's payoffs are not observable and therefore the equilibrium mixtures are not known to the observer. Nevertheless, the server's actions are observable, and equilibrium theory makes testable predictions about his behavior: it predicts that his expected payoff — the probability he will ultimately win the point being played — should be the same whether he serves to the receiver's forehand or to his backhand.

We have constructed a data set that contains detailed information about every point played in ten classic 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 forty point games. In each of the forty 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. These tests reject minimax play at the 5% level in only one of the forty point games; this rejection rate is actually slightly below the rate predicted by the random character of equilibrium play. We also conduct two joint tests, *i.e.*, tests whether the *distribution* of observed play across the forty point games is consistent with equilibrium theory. The test statistics' values are far from the rejection region in both of the joint tests.

Our tests for equality of the servers' winning probabilities are "limited information" tests, or indirect tests: since we do not have information about the actual equilibrium (minimax) mixtures, we cannot test directly whether a player is playing his minimax mixture, as one can typically do in experimental tests. However, when we apply our limited-information tests to the data from Barry O'Neill's celebrated experiment,[4] minimax play is rejected at the 5% level for ten of his fifty subjects (nearly the same rejection rate as in the direct tests by Brown & Rosenthal), and the joint tests reject minimax play at any reasonable level of significance.

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, again for both our tennis data and for O'Neill's experimental data, and both for individuals' play and for the joint distribution of play. Our tests indicate that the tennis players are not quite playing randomly: they switch their serves from left to right and vice versa somewhat too often to be consistent with random play. This is consistent with extensive experimental research in psychology and economics 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 O'Neill's data.

[4] O'Neill's ingenious experimental design avoided several weaknesses he had identified in others' prior attempts to test the theory. His subjects' observed play did indeed conform more closely to the theory than had the play of subjects in previous experiments (see, for example, Figure 1 in Erev & Roth). O'Neill's experiment thus provides the best vehicle for comparing experimental data with the observed play in our professional tennis matches.

The remainder of the paper is organized as follows: In Section 1 we briefly describe the relevant details of tennis and we present our simple 2×2 point game model. Section 2 identifies observable (testable) implications of equilibrium theory in our model, and Section 3 describes our data set. In Section 4 we present our statistical tests of the equality of winning probabilities by the players — the professional tennis players and, for a comparative benchmark, the subjects in O'Neill's experiment. In Section 5 we describe tests for random play, including serial independence. Section 6 contains some concluding remarks.

1 The 2×2 "Point Game"

We model each point in a tennis match as a simple 2×2 normal form game between two players.[5] 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.[6]

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 strokes by each player, or perhaps even after a second serve is played (if the first serve turns out to be a fault[7]). We do not attempt to model the play after the serve, but instead adopt a reduced-

[5] Essentially the same 2×2 model appears in Dixit & Nalebuff (1991).

[6] We consider alternative models of the point game in Section 2.

[7] If the serve fails to land in a designated region of the court, it is a "fault." A server is allowed one fault on each point — *i.e.*, if his first serve for the point is a fault, he gets to serve a second serve; if this second serve is also a fault, he loses that point.

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.[8] 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 which the reader may find helpful. The example's payoff numbers π_{sr} are hypothetical, but are numbers for which the example's properties are similar to the salient features of the data.)

The 2×2 point game depicted in Figure 1 can be used equally well to describe the strategic structure of first serves or second serves, but of course the probability-payoffs π_{sr} will typically be different for second serves than for first serves. Due to limitations in the data, which we describe in Section 3, we consider only first serves.

We have already suggested that the players have reason to be unpredictable in their actions. If the server, for example, knows (or is relatively certain) that the receiver is going to overplay to his left, then the server is more likely to win the point if he serves to the receiver's right than to his left. And if the server knows that the receiver is going to overplay to his right, he will have a better chance to win the point by serving to the left. In other words,

$$\pi_{LL} < \pi_{RL} \qquad \text{and} \qquad \pi_{RR} < \pi_{LR} \ . \tag{M$_1$}$$

Similarly, if the receiver knows the direction the server will choose, his return is likely

[8] 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.

to be more effective if he overplays in that direction than if he goes in the opposite direction; *i.e.*

$$1 - \pi_{LL} > 1 - \pi_{LR} \qquad \text{and} \qquad 1 - \pi_{RR} > 1 - \pi_{RL} \ ,$$

or equivalently,

$$\pi_{LL} < \pi_{LR} \qquad \text{and} \qquad \pi_{RR} < \pi_{RL} \ . \tag{M$_2$}$$

The four inequalities in (M$_1$) and (M$_2$) are together sufficient to ensure that the point game has a unique Nash equilibrium in which both players employ strictly mixed strategies. We therefore say that a point game satisfies the **Mixed Strategy Condition** if it satisfies (M$_1$) and (M$_2$). We assume that every point game in the tennis matches we are going to encounter satisfies the Mixed Strategy Condition:

Assumption 1: Every point in a tennis match is played as a 2×2 constant-sum normal form game that satisfies the Mixed Strategy Condition.

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. Indeed, an entire tennis match can also be modeled as an abstract game, in this case an infinite-horizon extensive form game comprised of point games: the winning of points in the point games determines the winning of *games* in the match;[9] the winning of games determines the winning of *sets*; and the winning of sets determines who wins the *match*. We say that a tennis match is an *infinite-horizon* game because of the scoring rules of tennis: a

[9]We unavoidably use the word "game" to mean two different things: (a) the traditional name for a unit of scoring in tennis, as in its italicized occurrence here; and (b) a game in the game-theoretic sense, an abstract model of a strategic situation. We use the term "abstract game" for the second notion, when it helps resolve ambiguity.

game does not end until one of the players is ahead by *two* points. (Moreover, a set played without the tiebreak rule does not end until one of the players is ahead by two games.[10]) Thus, there is no upper bound on the number of points that can be played in a match, even a match that uses the tiebreak rule in all sets.

The fact that the point games are merely the elements of a larger abstract 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 answer to this question — the link between the point games and the "match game" — is provided by the main result in Walker & Wooders (1999), where a class of games called *binary Markov games* is defined and analyzed.[11] The match game in tennis is an example of a binary Markov game. Thus, the Walker & Wooders result, when applied to tennis,[12] says that minimax (and therefore Nash equilibrium) play in the match game consists of the players playing, at every point in the match, a Nash equilibrium of the point game associated with that point, where the players' payoffs in the point game are indeed their respective probabilities of winning the point at hand, as in our Assumption 1 above. In other words, equilibrium play in a tennis match requires that a player play only to win the current point, ignoring the score (except to the extent that it directly affects the probability-payoffs π_{sr}), and ignoring the actions or outcomes on all previous points, as well as the effect the current point

[10] The longest such set in organized play continued for 96 games, ending in a score of 49-47. The same match included a 22-20 set as well.

[11] A Markov game is an extensive form game in which the current state is determined stochastically from just the state at the previous period and the players' actions at the previous period, and in which the transitions from state to state are otherwise independent of history. A Markov game is binary if from each state it is possible for the game to transit to at most two other states. Walker & Wooders show that in binary Markov games, a player's minimax behavior strategy requires him to play minimax for each point.

[12] See Walker & Wooders (1999), where the application to tennis is carried out.

will have on the remainder of the match.[13,14]

Variation in Point Games

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. 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.[15] Because of the players' particular abilities, the

[13]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." (Feinstein (1991).)

[14]It is sometimes said that players play differently — or that they *should* play differently — on "big points," that is, on points that are "more important." This might be true if, for example, a player is attempting to conserve his energy by focusing his attention on the most important points. We do not attempt to model such considerations here, and the theory in Walker & Wooders therefore implies that players should *not* (in equilibrium, or in minimax play) play any differently based on a point's importance.

[15]When the number of points completed so far in the current game is even, tennis rules require that the serve must land in the receiver's "deuce court" to be a "good" serve; otherwise it is a fault. And when the number of completed points is odd, the serve must land in the receiver's "ad court." Further, deuce-court points must be served from behind the deuce half of the server's baseline, and ad-court points from behind the ad half of the baseline. See Figure 2, which contains a diagram of the tennis court.

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.

2 On Testing the Theory

Our simple theoretical model of tennis makes some predictions about tennis players' behavior that we can subject to empirical testing. The theory's most obvious 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. His observed choices on first serves 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.

As usual in empirical research, however, several obstacles stand in the way of such

a straightforward approach. If we were constructing an experiment, we would have control over the players' equilibrium mixtures, because we could specify the entries in the game's payoff matrix. But in a tennis match the entries in the payoff matrix are not known and are not observable, at least not directly. There may appear to be some hope that we could *estimate* these probabilities from data that *are* observable. For example, if for each point in a match we could observe the two players' actual choices as well as the eventual winner of the point, then we could directly estimate the probability-payoffs π_{sr} for each of the four point games and use these to calculate our estimates of the players' equilibrium mixtures in each of the point games. But for all practical purposes, the receiver's choices are simply unobservable.

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. Fortunately, the theory *does* make some predictions about these observable elements of the game. The prediction on which we will focus most of our attention concerns the server's **win rates** for each of his alternative actions. According to Assumption 1, each point game in a given match has a unique equilibrium, and each player uses a strictly mixed strategy in the equilibrium. Thus, in a given point game, if the players are playing according to the equilibrium then each player's expected payoff from playing Left must be the same as his expected payoff from playing Right. For each of the players in a match, this prediction *for his serves* can be confronted with data: for every point that he serves, we can observe the action he chose and the eventual winner of the point. Hence, we can formulate and test the hypothesis that his expected payoffs are the same for Left and Right serves.

Alternative Models of the Point Game

Our analysis of the point game has been simplified by assuming that the server and the receiver each have available only two actions at each stage. We now argue that the equality of winning probabilities for serves to the Left and serves to the Right remains

an implication of equilibrium theory under alternative, more detailed models of the point game. There are several issues here. First, it is possible to deliver a serve which is neither to the receiver's left nor to his right — indeed, players do occasionally serve intentionally directly toward the receiver's body (i.e., to the Center). Note, however, that in a mixed-strategy Nash equilibrium, if serves to the left, right, and center are each made with positive mixture probability, then the expected payoff (i.e., the probability of winning the point) must be the same for each of the directions. In particular, the probability of winning the point must be the same for Left as for Right, as in our hypothesis. Serves to the Center are infrequent in our data, and we therefore ignore them for statistical purposes.

A second issue is that it is possible to vary the serve in more dimensions than simply its direction. One can deliver a flat serve, a kick serve, a slice serve, and so on, and each of these can be delivered at various speeds. We ignore these distinctions for several reasons: for example, they are difficult for an observer to distinguish; and if we *were* to distinguish them, the resulting large number of strategies would leave us with small numbers of observations in each category, and thus with less powerful statistical tests: we would be less likely to reject the theory if it is false. Effectively, then, we are pooling various kinds of serve to the left into just one strategy, Left, and similarly for Right. Again, note that in a mixed-strategy Nash equilibrium, the probability the server will win the point, conditional on the serve being sliced-and-to-the-left, must be equal to the probability he will win the point conditional on the serve being flat-and-to-the-left (if both receive positive mixture probability). This probability, in turn, is the same as the probability he will win the point conditional only on the serve being to the Left. Hence, the hypothesis of equality of winning probabilities holds if play is governed by the Nash equilibrium of the point game, even if some types of serves are pooled, just as it does if some types of serves are omitted.

A third issue is that the Receiver may not actually overplay to one side or the

other. An alternative model of the point game can be developed in which the Receiver has a continuum of actions available to him, representing locations across the baseline where he can position himself to await the serve. Under plausible assumptions, this alternative model also has a unique equilibrium, in which each player is playing his minimax strategy and in which the Server's winning probabilities are the same when he serves to the Right as when he serves to the Left. We have used the more parsimonious 2×2 model for its simpler, more easily understood structure. Equilibrium theory's predictions in the 2×2 game are identical to its predictions in this alternative game.

The point game can also be modeled as an extensive form game, to capture the temporal and contingent structure of a point. The players make their Left-or-Right choices at the first serve; then nature chooses whether the serve is good or is a fault; if the serve is good, nature chooses which player wins the point, and if the serve is a fault the players make another Left-or-Right choice at the second serve, and nature again chooses whether the serve is good and if so which player wins the point — in each case with a new set of "second-serve" probabilities. It is straightforward to reduce such an extensive form game to the 2×2 point game we introduced in the preceding section.

3 The Data

Our data set was obtained from videotapes of classic 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 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, 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).

We were able to obtain videotape of ten such matches, and our data set is comprised of these ten matches. Recall that every tennis match contains four point games, so we have data for forty 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 (and, in a moment, Table 2) as the data from an "experiment" for which we model the data generating process as a 2×2 point game, as in Section 1. 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. These data are summarized in Table 1. The columns labelled Serve Direction 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), right (R), or center (C). The columns labelled Points Won indicate, for each direction of first serve, the number of times the server ultimately won the point.[16] For example, the first row of Table 1 reports the serve and win data for Ken Rosewall from the ad court in the 1974 Wimbledon semifinal (in which he

[16] 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.

defeated Stan Smith). In that match Rosewall won 25 of the 37 first serves to the left, he won 26 of the 37 first serves to the right, and he won 2 of the 4 first serves to the center. 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 second column of Table 1, labelled Hand, indicates whether the player is right or left handed. 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. Only 6% of first serves were to the center, and we therefore ignore center serves: accounting for them would have only a negligible effect on our results.

4 Testing for Equality of Winning Probabilities

In any point game which satisfies the Mixed Strategy Condition, the theory of mixed-strategy equilibrium holds that the server's winning probabilities will be the same when serving to the receiver's Left as when serving to his Right. In this section we formulate this theoretical implication as a *null hypothesis*, and we test the hypothesis for the matches in our data set. We find that the data do not come close to rejecting the null hypothesis. Indeed, the observed distribution of the test statistic's realized values in the forty "experiments" looks remarkably similar to the distribution predicted by the theory.

In order to provide a comparison with experimental data, we also conduct the same tests on the data from O'Neill's experiment, in which subjects' play adhered more closely to the minimax mixtures than in other experiments. The tests yield overwhelming rejection of the theoretical prediction, in contrast with the results when the same tests are applied to the tennis data.

Hypotheses and Tests for the Tennis Data

We first test, for each of the forty 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 forty 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 Pearson's chi-square goodness-of-fit test of equality of two distributions (see, for example, p. 449 of Mood, Graybill, and Boes (1974)). We index the forty point-game experiments by i ($i = 1, \ldots, 40$). For experiment i, let n_j^i denote the number of first serves that were delivered in direction $j \in \{L, R\}$, *i.e.*, the number of points that began with serves in direction j. Each such point was ultimately won by either the server or the receiver; we say that the outcome was either a *success* (if the server won the point) or a *failure* (if the receiver won the point), which we denote by S and F. Let N_{jS}^i and N_{jF}^i denote the number of first serves in direction j for which the ultimate outcome was S or F, respectively. Let p_j^i denote the (true, but unknown) probability that the server will win the point (*i.e.*, the probability of a success) when the first serve is in direction j. For each experiment i, then, our **null hypothesis** is that $p_L^i = p_R^i$, or equivalently, that there is a number p^i such that

$p^i_L = p^i$ and $p^i_R = p^i$.[17] If the null hypothesis is true, then the Pearson statistic

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

is distributed asymptotically as chi-square with 2 degrees of freedom. Since p^i is unknown, it must be estimated. Under the null hypothesis, the maximum likelihood estimate of p^i is $\frac{N^i_{LS}+N^i_{RS}}{n^i_L+n^i_R}$; we therefore replace p^i in the expression (1) for Q^i with this estimate. The asymptotic distribution of this Pearson test statistic Q^i is chi-square with 1 degree of freedom.

Table 2 reports the results of the Pearson test. For each of the forty point-game experiments, the two columns labelled "Pearson statistic" and "p-value," at the righthand side of the table, report the value of the test statistic Q^i along with its associated p-value (*i.e.*, the probability that a draw from the chi-square-1 distribution will be at least as large as the observed value of the test statistic). The null hypothesis is rejected, for example, at the 5% significance level if the p-value is .05 or smaller. In only one of our forty point-game experiments (Sampras serving to Agassi in the deuce court in 1995) do we find the null hypothesis rejected at the 5% level, and for only one other point game (Connors serving to McEnroe in the ad court in 1982) do we reject at the 10% level. Note that with forty 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% level and four at the 10% level. Considering simply the number of 5% and 10% rejections, then, the tennis data appear quite consistent with the theory.

This suggests a test of the *joint* hypothesis that the data from *all forty* experiments were generated by equilibrium play. We apply Pearson's test to the joint hypothesis that $p^i_L = p^i_R$ for *each one* of the experiments $i = 1, \ldots, 40$ (but allowing

[17] The null hypothesis is an implication of Assumptions 1 and 2 and the assumption that the players are playing the unique equilibrium of the point game. These assumptions yield an additional implication, which is also a part of our null hypothesis: that both the Left and Right processes are *i.i.d.* and that they are independent of one another.

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 forty individual tests we have just described, *i.e.*, $\sum_{i=1}^{40} Q^i$, 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 forty 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 forty 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 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 forty Q^i values in Table 2 should be 40 such chi-square draws. Equivalently, the p-values associated with the realized Q^i values (also in Table 2) should have been forty draws from the uniform distribution $U[0,1]$.[18]

A simple visual comparison of the observed distribution with the theoretically predicted distribution is provided in Figure 3, a histogram of the forty observed p-values. The horizontal axis contains the ten deciles of possible values, 0 to .10, .10 to .20, *etc.*; the height of a decile's bar indicates the number of p-values that fell in that decile according to the p-value column of Table 2. The horizontal line indicates the height of a uniform histogram with 40 observations. This informal visual comparison

[18] To see that the p-values (for any c.d.f.) are distributed $U[0,1]$ under the null hypothesis, suppose that a random variable X has c.d.f. $F(x)$. If the realized value of X is x, then the associated p-value is $p(x) = 1 - F(x)$. For any b satisfying $0 \leq b \leq 1$, we have $\Pr[p(X) \leq b] = \Pr[1 - F(X) \leq b] = \Pr[F(X) \geq 1 - b] = \Pr[X \geq F^{-1}(1-b)] = 1 - F(F^{-1}(1-b)) = 1 - (1-b) = b$.

suggests that the data are consistent with the theory: the empirical histogram is not dramatically different than the theoretical uniform histogram.

A more detailed comparison is provided by Figure 4, in which the empirical cumulative distribution function (the *c.d.f.*) of the forty 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 4. The empirical and the theoretically predicted distributions depicted in Figure 4 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 4, the KS test is also more powerful than the Pearson joint test against many alternative hypotheses about how the data were generated.[19]

Formally, the KS test for the tennis data is described as follows. The hypothesized c.d.f. for the p-values is the uniform distribution, $F(x) = x$ for $x \in [0, 1]$. The empirical distribution of the 40 p-values in Table 2, denoted $\hat{F}(x)$, is given by $\hat{F}(x) = \frac{1}{40}\sum_{i=1}^{40} I_{[0,x]}(p^i)$, where $I_{[0,x]}(p^i) = 1$ if $p^i \leq x$ and $I_{[0,x]}(p^i) = 0$ otherwise. Under the null hypothesis, the test statistic $K = \sqrt{40}\sup_{x\in[0,1]} |\hat{F}(x) - x|$ has a known distribution (see p. 509 of Mood, Boes, and Graybill (1974)). For the tennis data in Table 2, we have $K = .670$, with a p-value of .76, far toward the opposite end of

[19] For a simple illustration of this, consider an example in which 40 subjects are instructed to each toss a fair coin 100 times and to record the number of tosses that are heads. Suppose each subject reports that he obtained exactly 50 heads, and we wish to test the joint hypothesis that each subject tossed the coin "randomly" to generate his report. The value of each of the Pearson statistics would be zero, and therefore their sum would be zero, and therefore we would not reject the null hypothesis at any level of significance using the Pearson test. But the KS test would decisively reject the joint null hypothesis: the empirical c.d.f. of the 40 realized values of the Pearson statistic (*viz.*, a unit mass on the value zero) is not "close" to the distribution of Pearson values under the null hypothesis (*viz.*, the chi-square-1 distribution).

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% of the time, and a "better" (smaller) value of K only 24% of the time.

Statistical analysis thus indicates convincingly that the empirical winning probabilities in the tennis data are consistent with the theory of mixed-strategy equilibrium. However, by itself this might not be convincing evidence that the players are indeed playing their minimax (equilibrium) mixtures: our tests might have little power to detect deviations from minimax play, because they use observed win rates to test hypotheses about equality of winning probabilities, rather than testing directly whether the players' observed mixtures could have been generated by play according to their equilibrium mixtures. (Recall that since the exact payoffs in the normal form point games are not observable, the equilibrium mixtures are also unknown.)

In order to evaluate this possibility, and to provide a comparison of the tennis data with experimental data, we first apply the same tests to the data from Barry O'Neill's (1987) experiment, the experiment in which observed play adhered most closely to the equilibrium prediction, and then we conduct a formal analysis of the power of the Pearson joint test against a range of alternative hypotheses.

Applying our Tests to O'Neill's Data

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. Each play of the game is won by one player or the other. The Row player wins if both players choose the Joker or if the players choose different number cards. Otherwise the Column player wins. The game has a unique Nash equilibrium: each player chooses the Joker with probability .4 and chooses each number card with probability .2. O'Neill's subjects all played the game 105 times, each subject always playing against the same opponent. 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 fifty 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 "limited information" tests as we have already carried out above for the tennis data.

Table 3 presents the results of the Pearson goodness-of-fit tests. The first two columns of the table identify the pairs and the players. The subjects' observed mixtures are shown on the left-hand side of the table, and the results of the limited-information Pearson tests on the subjects' win rates appear on the right-hand side. In the fifty 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% level and 15 rejections at the 10% level. Figure 5 depicts a histogram of the fifty p-values.[20]

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 (the associated p-value is 1.239×10^{-14}), and 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 6 is the analogue for O'Neill's data of Figure 4 for the tennis data: it depicts the empirical distribution and the hypothesized (*i.e.*, uniform) distribution of

[20]Brown & 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.

the p-values for the tests of equality of winning probabilities in Table 3. For O'Neill's data the value of the KS test statistic is $K = 1.704$, with a p-value of .006. Hence the KS test rejects the null hypothesis at significance levels as low as 1%. Comparison of Figures 3 and 4 with Figures 5 and 6 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.

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.[21] Using the point game in Figure 1, we formulate a parametric class of plausible alternative hypotheses and we conduct Monte Carlo simulations to evaluate the power of the Pearson test to reject the null hypothesis when any of these alternative hypotheses is true. As we have already indicated, the point game 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 $.53\frac{1}{3}$, while in the data 53.5% 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 .65, while in the data the servers won 64.7% of all points. 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. We streamline the analysis by using the "typical" point game of Figure 1.

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

[21] We omit discussion of the power of the KS joint test of equality of winning probabilities. Our Monte Carlo simulations reveal that the KS test has less power than the Pearson test for the class of alternative hypotheses we consider here. Moreover, for this class of alternatives the KS test rarely rejects the null hypothesis unless the Pearson test also rejects.

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

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

and

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

Under the joint null hypothesis that the receiver in each experiment follows his equilibrium mixture (and that the equilibrium mixture prescribes Left two-thirds of the time, as in the "typical" point game) — *i.e.*, that $\theta = \frac{2}{3}$ in each experiment — the Pearson test statistic $\sum_{i=1}^{40} Q^i$ is asymptotically distributed as chi-square with 40 degrees of freedom.

At the 5% significance level, the Pearson joint test consists of rejecting the null hypothesis if $\sum_{i=1}^{40} Q^i$ 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 $\sum_{i=1}^{40} Q^i$ is not known. We have used Monte Carlo methods to estimate the power of the test against alternative values of θ.[22] The power function is depicted in Figure 7. One can see that the Pearson joint test has significant ability to reject the null

[22]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 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. (In each experiment there were 40 serves in direction L and 35 serves in direction R; in the data the average number of first serves is 40.55 to the Left and 35.10 to the Right.)

hypothesis when the true value of θ differs significantly from $\theta_0 = \frac{2}{3}$. For example, if $\theta = .5$ — *i.e.*, if the receivers actually choose Left and Right with equal probability — the joint null hypothesis is rejected with probability .58. Thus, the test has fairly high power, even though the server's winning probabilities are not very different (they are $p_L = .685$ and $p_R = .610$). Under the alternative hypothesis that $\theta = .6$, *i.e.*, that the receiver chooses Left with probability .6, the power of the test rises dramatically to .98.

5 Serial Independence

Our Assumptions 1 and 2 require that a player's mixed strategy be the same at each occurrence of a point game — *i.e.*, each time a particular player serves from a particular court. This implies that the server's choices at each occurrence of a given point game will be serially independent — that his serves from a given court are generated by a single *i.i.d.* binomial process. In this section we test whether the first serves by the players in our data set were indeed serially independent. We also provide a comparison with O'Neill's experimental data by testing for serial correlation in his subjects' play and comparing the results to our tests for serial correlation in the tennis data.

Hypotheses and Tests for the Tennis Data

First we test individually, for each of the forty "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, *i.e.*, a string which

is not part of any longer string of identical symbols.[23]) 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 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 Gibbons and 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.*, $F(r; n_L, n_R) = \sum_{k=1}^{r} f(k; n_L, n_R)$, the probability of obtaining r or fewer runs. At the 5% significance level, the null hypothesis of serial independence in experiment i is rejected if either $F(r^i; n_L^i, n_R^i) < .025$ or $1 - F(r^i - 1; n_L^i, n_R^i) < .025$, *i.e.*, if the probability of r^i or fewer runs is less than .025 or the probability of r^i or more runs is less than .025.

Table 4 shows the data and the results for our tests of serial independence. For each of the forty 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% significance level, the null hypothesis is rejected in five of the forty experiments (the expected number of 5% rejections is only two). In three cases the null hypothesis is rejected because there are too many runs, and in the other 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 employ the Kolmogorov-Smirnov goodness-of-fit test. The KS

[23]For example, the sequence $s = (L, L, R, L)$ has three runs. We omit serves to the center.

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 4. Under the null hypothesis of serial independence in experiment i, the statistic t^i is distributed $U[0,1]$.[24]

The empirical c.d.f. of the realized values $t^1,\ldots,t^{40}$ in Table 4 is depicted in Figure 8. The value of the Kolmogorov-Smirnov test statistic is $K=1.948$,[25] with a p-value of .001. Hence, we reject the null hypothesis that in all forty experiments the first serves were serially independent. Figure 8 and Table 4 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 of actions they generally "switch too often" to be consistent with randomly generated choices (Wagenaar, 1972).

Serial Independence in O'Neill's Data

Table 5 shows the data and the results for tests of serial independence in O'Neill's

[24] Let $x\in[0,1]$, let n_L and n_R be fixed, and write $F(r)$ for $F(r;n_L,n_R)$. We show that under the null hypothesis $\Pr[t\leq x]=x$, — *i.e.*, t is uniformly distributed. Let $\bar{r}$ be such that $F(\bar{r}-1)\leq x<F(\bar{r})$. We have $\Pr[t\leq x]=f(2)+\ldots+f(\bar{r}-1)+f(\bar{r})\frac{x-F(\bar{r}-1)}{F(\bar{r})-F(\bar{r}-1)}=x$, because $F(\bar{r})-F(\bar{r}-1)=f(\bar{r})$ and $f(2)+\ldots+f(\bar{r}-1)=F(\bar{r}-1)$.

[25] This "randomized" test was performed many times. While there was of course variation in the forty 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.

experiment. As before, we pool the three strategically equivalent number cards in his game into a single choice, "non-Joker," and we distinguish only between Jokers and non-Jokers when counting runs. For each of O'Neill's fifty subjects, the columns J and N 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% 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 5 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 9. The value of the Kolmogorov-Smirnov test statistic is $K = 2.503$, with a p-value of .000007.[26] Hence we reject the joint null hypothesis that each of the fifty 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 8 and 9 suggests that while play is negatively correlated in both the tennis data and O'Neill's experimental data (generally too much switching between choices), 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.

[26] Again, many trials produced little variation in the value of K or the p-value.

6 Concluding Remarks

The theory of mixed strategy equilibrium has not been consistent with the empirical evidence gathered through more than forty years of experiments involving human subjects. Our objective here has been to identify, if possible, *some* class of strategic encounters in which people behave as the theory predicts. We have suggested that important matches between the world's best tennis players provide a class of situations which are not only theoretically and empirically tractable, but which also differ dramatically from experimental encounters in some crucial respects: top tennis players have devoted their lives to becoming expert at their games; they know the physical and strategic characteristics of one another's play; and their financial motivation to exploit any strategic opportunity is extraordinarily large.

Beginning with a simple 2×2 game-theoretical model of players' actions on first serves, and with a data set compiled from classic professional tennis matches, our theoretical and empirical analysis has shown that the serve-and-return play of the tennis players in our data set is remarkably consistent with the theory of mixed strategy equilibrium in every respect but one: the players exhibit a tendency to switch the direction of their serves from left to right, or vice versa, somewhat too often to be truly random. The contrast between the consistency of our data with the theory, and the failure of experimental data to conform to the theory, is striking.

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 in strategic situations requiring unpredictable play, inexperienced players will not generally mix in the equilibrium proportions. Similarly, we do not mean to suggest that the theory applies only to people who have developed years of experience in the particular strategic context at hand. Rather, we have identified a characteristic of strategic encounters — *viz.*, the participants' experience and expertise with the strategic situation — that may be an important factor in determining whether equilibrium theory, in its current form, provides a useful model of behavior.

We have provided the first empirical evidence that the theory of mixed strategy equilibrium may indeed be an empirically useful theory. There is a spectrum of experience and expertise, with novices (such as typical experimental subjects) at one extreme and our 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 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 the behavior of inexperienced players.

References

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

[2] Dixit, A. and B. Nalebuff (1991). *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life*, New York: W.W. Norton.

[3] Erev, I. and A. Roth (1998). "Modeling How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *American Economic Review* **88**, 848-881.

[4] Feinstein, John (1991). *Hard Courts*, New York: Random House, Inc.

[5] Gibbons, J. and S. Chakraborti (1992). *Nonparametric Statistical Inference*, New York: Marcel Dekker, Inc.

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

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

[8] Wagenaar, W. (1972). "Generation of Random Sequences by Human Subjects: A Critical Survey of the Literature," *Psychological Bulletin* **77**, 65-72.

[9] Walker, M. and J. Wooders (1999). "Binary Markov Games," University of Arizona, photocopy.