SIAM Review
Vol. 48, No. 4, pp. 722–742

Monte Carlo Tennis*

Paul K. Newton[†]
Kamran Aslam[†]

Abstract. The probability of winning a game, set, match, or single elimination tournament in tennis is computed using Monte Carlo simulations based on each player's probability of winning a point on serve, which can be held constant or varied from point to point, game to game, or match to match. The theory, described in Newton and Keller [*Stud. Appl. Math.*, 114 (2005), pp. 241–269], is based on the assumption that points in tennis are independent, identically distributed (i.i.d.) random variables. This is used as a baseline to compare with the simulations, which under similar circumstances are shown to converge quickly to the analytical curves in accordance with the weak law of large numbers. The concept of the *importance* of a point, game, and set to winning a match is described based on conditional probabilities and is used as a starting point to model non-i.i.d. effects, allowing each player to vary, from point to point, his or her probability of winning on serve. Several non-i.i.d. models are investigated, including the "hot-hand-effect," in which we increase each player's probability of winning a point on serve on the next point after a point is won. The "back-to-the-wall" effect is modeled by increasing each player's probability of winning a point on serve on the next point after a point is lost. In all cases, we find that the results provided by the theoretical curves based on the i.i.d. assumption are remarkably robust and accurate, even when relatively strong non-i.i.d. effects are introduced. We end by showing examples of tournament predictions from the 2002 men's and women's U.S. Open draws based on the Monte Carlo simulations. We also describe Arrow's impossibility theorem and discuss its relevance with regard to sports ranking systems, and we argue for the development of probability-based ranking systems as a way to soften its consequences.

Key words. tennis, Monte Carlo Method, non-i.i.d. effects, probabilistic ranking systems, Arrow's theorem

AMS subject classifications. 65C05, 91A60, 60J20, 91B12, 65Q05

DOI. 10.1137/050640278

1. Introduction.

We describe a Monte Carlo method which we use to calculate the probability of winning a game, set, match, or single elimination tournament in tennis. The corresponding analytical theory, described in detail in Newton and Keller [15] and hereafter referred to as Part I, is based on each player's probability of winning a point on serve. Thus the values $p_A^R \in [0,1]$ and $p_B^R \in [0,1]$ for player A and player B are assumed constant throughout each match and throughout the tournament. This is the assumption that points in tennis are independent, identically distributed (i.i.d.) random variables. In practice, these values are obtained empirically from each of the player's statistics gathered over enough matches and against different opponents so that this accumulated value of the ratio of points won on serve over total points served for each player can be used predictively. This ratio converges quite rapidly to

[*] Received by the editors September 13, 2005; accepted for publication (in revised form) March 2, 2006; published electronically November 2, 2006. This work was supported in part by the National Science Foundation under grants NSF-DMS 9800797 and NSF-DMS 0203581.

http://www.siam.org/journals/sirev/48-4/64027.html

[†] Department of Aerospace & Mechanical Engineering and Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1191 (newton@spock.usc.edu).

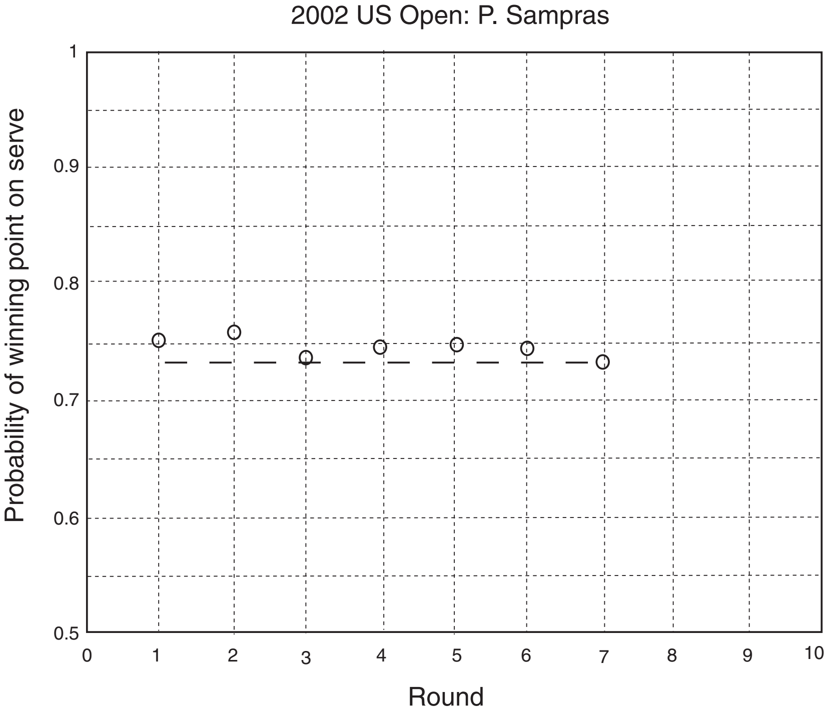


Fig. 1 *Convergence of* p_A^R *to the sample mean* $\mu = 0.73$ *for Pete Sampras through the seven rounds of the* 2002 *U.S. Open. Match mean* $\mu_m = 0.7392$ *with standard deviation* $\sigma = .0314$*.*

a nearly constant value for each player, as shown, for example, in Figure 1, which shows data for Pete Sampras, the winner of the 2002 U.S. Open men's singles event, his last tournament win before retirement. Each data point represents the ratio of total points won on serve over total points served accumulated through the first n rounds of the tournament. The final data point ($n = 7$) contains information on the player accumulated through the entire tournament, and hence can be viewed as a cumulative value over his or her field of opponents. As an example, this ratio for Pete Sampras converges fairly rapidly to its cumulative value of 0.73, which would be used as input for him in the analytical theory described in Part I.

Using this value for Pete Sampras (i.e., $p_A^R = 0.73$), we can ask what the probability of defeating him would be, given the full range of values for p_B^R. Figure 2 shows the results based on the analytical theory from Part I. The curves depict the probabilities of winning a game, set, and match against Sampras, along with data from the 2002 U.S. Open. A general conclusion based on the steepness of the analytical curves throughout the typical range encountered on the professional circuit ($0.60 < p_A^R, p_B^R < 0.75$) together with the fact that the curves for sets are steeper than those for games, and are steeper yet for matches (see the corresponding curves in Figure 2), is that *the better player usually wins in tennis*—the scoring system conspires against the weaker player. This is depicted dramatically by the data points shown in the figure. The data, taken from his 2002 U.S. Open opponents, show that by winning 65% of their points on serve, they won over 80% of their service games, but only 17% of the sets and none of their matches. Relatively small differences in ability between

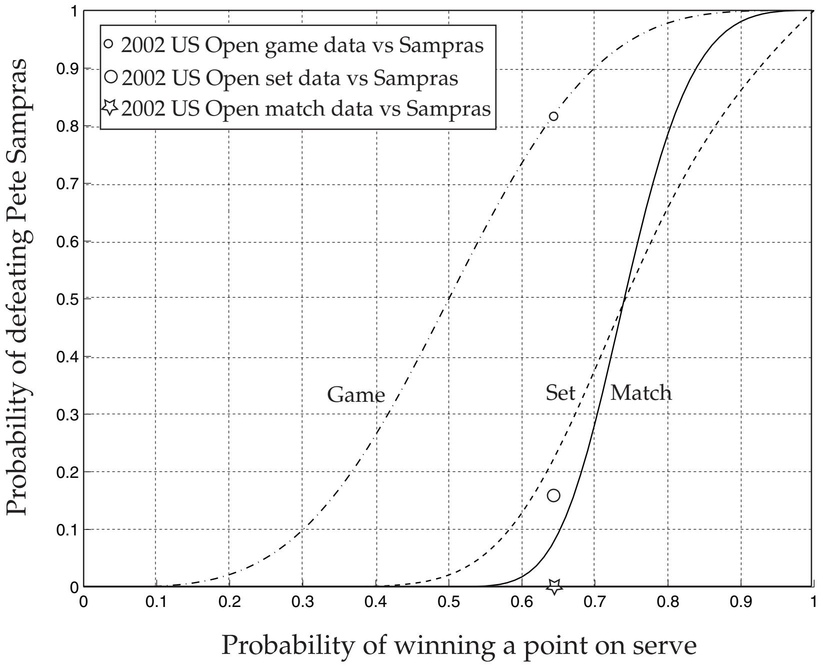


Fig. 2 *Probability of winning against Pete Sampras in the* 2002 *U.S. Open tournament. Note that the steepness of the curves increases in going from games to sets to matches.*

players are amplified relentlessly against the weaker player by the way the scoring system is constructed. In addition, because of the fact that the top players are spread throughout the draw based on the seeding system, they tend to meet often (typically in the semifinals or finals) during a season of tournaments. These facts should, in principle, make ranking systems for tennis easier to construct than in other sports which have a more random component, making upsets more common, and have top ranked teams that may never play each other during a season (see discussions in [4] and [6]).

There are several important points to make. First, despite the fact that the convergence shown in Figure 1 is rapid, there are always fluctuations of the higher-order moments around the mean. These fluctuations are typically small (roughly 1%) compared to the difference with the average of other players. Hence, on a match-by-match basis, each player's ratio of points won on serve to points served varies somewhat from his accumulated value gathered over large numbers of matches against different opponents. For example, in the case of Pete Sampras for the 2002 U.S. Open, his match mean (i.e., the mean of the 7 ratios associated with each match) was 0.7392 with a standard deviation of .0314. For women, the match mean tends to be lower but the variation tends to be higher. The match mean associated with Serena Williams in the 2002 U.S. Open and Wimbledon championships, both of which she won, was 0.7158 with a standard deviation of .0762.

In practice, this means that for any given match, even with a large amount of data in hand, there is some uncertainty as to what actual values to take for p_A^R and p_B^R for each of the players. In addition, when examining targeted homogeneous data sets, such as, for example, the Borg–McEnroe series of head-to-head matches, there is some evidence of non-i.i.d. effects creeping in, such as the so-called back-to-the-wall and hot-hand effects (see [7, 10] for more discussions). Other aspects of tennis that potentially introduce non-i.i.d. effects are the introduction of new balls [12] or psychological factors that could be present in the first or final set of a match [11, 13]. Although we think of these kinds of effects as "second order," in close matches they can play prominent roles because of the steepness of the curves in Figure 2.

One can then legitimately ask how accurate the analytical theory is in predicting probabilities of winning *individual* matches, given that it uses fixed values of p_A^R and p_B^R throughout each match. In principle, one could compare the theory with data gathered from the tournaments; however, this is difficult. While the number of points and games played by each player in a typical match and tournament is large enough to extract meaningful statistics (as shown in Part I when comparing the theory with data for points and games), the number of sets and matches played is not. What is worse, the analytical formulas predicting set and match probabilities are functions of both p_A^R *and* p_B^R, while the formulas for predicting tournament outcomes are functions of 128 variables, one for each of the players in the tournament. Hence, gathering data for an individual player, say, player A, requires that one look at this player's matches only against opponents with the same value of p_B^R, which reduces the data set even further. One might then be tempted to look at data for each player over an entire season of tournaments against opponents with one value of p_B^R. But this introduces other problems as tournaments are played on several different surfaces and player characteristics can vary widely from surface to surface. For example, although Pete Sampras was dominant on grass, winning a record seven Wimbledon singles titles, he never did well at the French Open championship, which is played on the much slower clay surface.

For all of these reasons, it seems desirable to develop a Monte Carlo approach that is capable of generating large data sets quickly and reliably that would be difficult, if not impossible, to gather in practice, and that could be used to evaluate the robustness of the i.i.d. assumption adopted in Part I and used in other analytical approaches that predate this work, such as those of Carter and Crews [3] and Pollard [17]. As discussed in the final section, such a simulation could also be used as the basis for a probabilistic ranking system that would have certain advantages over current ranking systems (see [4, 5] for a description of a random walk method for ranking football teams).

The search for evidence of non-i.i.d. effects has been pursued in several sports with mixed results. In basketball, an investigation into whether or not points are i.i.d. was pioneered by Tversky and Gilovich [24]. Their analysis of consecutive shots showed that, contrary to popular belief, the chances of a player hitting a shot are as good after a miss as after a hit; thus they found no evidence of a hot-hand effect. A similar analysis of hitting streaks in baseball [1, 23] also failed to detect any significant effects on the probability of making a hit due to a player's recent history of success or failure. In tennis, the question of whether points are i.i.d. random variables was first addressed in the paper of Klaassen and Magnus [10] by doing a statistical analysis of 90,000 points at Wimbledon, collected over a wide range of matches. This analysis *does* show some evidence that winning the previous point has a positive effect on winning the current point and that "important" points are more difficult to win for the server than points that are less important. Their ultimate conclusion, however, was that although points in tennis are not exactly i.i.d. random variables, the deviation from i.i.d. is small. Recent attempts to model some of these non-i.i.d. effects in tennis can be found in [7, 16, 18, 19, 20, 21].

The primary goal of our paper is to introduce a Monte Carlo approach in which we can investigate the effects of these deviations and to explore the effects of some specific non-i.i.d. models. We end with a description of Arrow's impossibility theorem and its consequences with regard to "outcome-based" (deterministic) ranking systems and give a brief description of a "predictive" ranking system (probabilistic) for tennis which could be used as the basis for an improved method for seeding players in a tournament, ultimately producing a year-end ranking for the ATP (Association of Tennis Professionals) and WTA (Women's Tennis Association) tours.

1.1. Analytical Theory. First, we review briefly the theory presented in Part I. In order to calculate the probability that one player A wins a tennis match against another player B, it suffices to know the probability p_A^R that A wins a rally when A serves, and the probability p_B^R that B wins a rally when B serves. When these two independent parameters are held constant throughout the match, explicit formulas for the probabilities of winning a game, set, and match for each player can be obtained. For example, the probability of winning a game on serve, p_A^G, is given by

$$p_A^G = (p_A^R)^4[1 + 4q_A^R + 10(q_A^R)^2] + 20(p_A^R q_A^R)^3 (p_A^R)^2 [1 - 2p_A^R q_A^R]^{-1}, \tag{1.1}$$

where $q_A^R = 1 - p_A^R$. Note that the formula depends only on characteristics of player A and not on player B. This simple, explicit, and compact formula which encodes nicely the game scoring system was, to our knowledge, first obtained in Carter and Crews [3]. As mentioned earlier, its relatively steep slope in the region of interest for most players ($0.55 \leq p_A^R \leq 0.75$) is responsible for the amplification of small differences in abilities, making upsets more rare in tennis than in other sports, such as football and basketball.

To obtain corresponding formulas for the probability of winning a set and a match, let p_A^S denote the probability that player A wins a set against player B, with A serving first, and $q_A^S = 1 - p_A^S$. To calculate p_A^S in terms of p_A^G and p_B^G, we define $p_A^S(i,j)$ as the probability that, in a set, the score becomes i games for A, j games for B, with A serving initially. Then

$$p_A^S = \sum_{j=0}^{4} p_A^S(6,j) + p_A^S(7,5) + p_A^S(6,6) p_A^T. \tag{1.2}$$

Here, p_A^T is the probability that A wins a 13-point tiebreaker with A serving initially, and $q_A^T = 1 - p_A^T$.

To calculate the terms $p_A^S(i,j)$ needed in (1.2), we solve the following system of recursion equations:

For $0 \leq i, j \leq 6$,
if $i - 1 + j$ is even: $p_A^S(i,j) = p_A^S(i-1,j) p_A^G + p_A^S(i,j-1) q_A^G$,
omit $i - 1$ term if $j = 6$, $i \leq 5$;
omit $j - 1$ term if $i = 6$, $j \leq 5$;

if $i - 1 + j$ is odd: $p_A^S(i,j) = p_A^S(i-1,j) q_B^G + p_A^S(i,j-1) p_B^G$,
omit $i - 1$ term if $j = 6$, $i \leq 5$;
omit $j - 1$ term if $i = 6$, $j \leq 5$.

These must be supplemented with the initial conditions

$$p_A^S(0,0) = 1, \quad p_A^S(i,j) = 0 \tag{1.3}$$

if $i < 0$ or $j < 0$. In terms of $p_A^S(6,5)$ and $p_A^S(5,6)$, we have

$$p_A^S(7,5) = p_A^S(6,5) q_B^G, \quad p_A^S(5,7) = p_A^S(5,6) p_B^G. \tag{1.4}$$

To calculate the probability of winning a tiebreaker, p_A^T, in terms of p_A^R and p_B^R, we define $p_A^T(i,j)$ to be the probability that the score becomes i for A, j for B in a tiebreaker with A serving initially. Then

$$p_A^T = \sum_{j=0}^{5} p_A^T(7,j) + p_A^T(6,6) \sum_{n=0}^{\infty} p_A^T(n+2,n). \tag{1.5}$$

Because the sequence of serves in a tiebreaker is A, BB, AA, BB, etc., we have

$$\begin{aligned} p_A^T(n+2,n) &= \sum_{j=0}^{n} (p_A^R p_B^R)^j \left(q_A^R q_B^R\right)^{n-j} \ \frac{n!}{j!(n-j)!} p_A^R q_B^R \\ &= (p_A^R p_B^R + q_A^R q_B^R)^n p_A^R q_B^R. \end{aligned} \tag{1.6}$$

Using (1.6) in (1.5) and summing yields

$$p_A^T = \sum_{j=0}^{5} p_A^T(7,j) + p_A^T(6,6) p_A^R q_B^R \left[1 - p_A^R p_B^R - q_A^R q_B^R\right]^{-1}. \tag{1.7}$$

To calculate $p_A^T(i,j)$, we solve the following equations:
For $0 \le i, j \le 7$,
if $i - 1 + j = 0, 3, 4, \dots, 4n-1, 4n, \dots$,

$$p_A^T(i,j) = p_A^T(i-1,j) p_A^R + p_A^T(i,j-1) q_A^R, \tag{1.8}$$

omit $j-1$ term if $i = 7$, $j \le 6$;
omit $i-1$ term if $j = 7$, $i \le 6$;

if $i - 1 + j = 1, 2, 5, 6, \dots, 4n+1, 4n+2, \dots$,

$$p_A^T(i,j) = p_A^T(i-1,j) q_B^R + p_A^T(i,j-1) p_B^R, \tag{1.9}$$

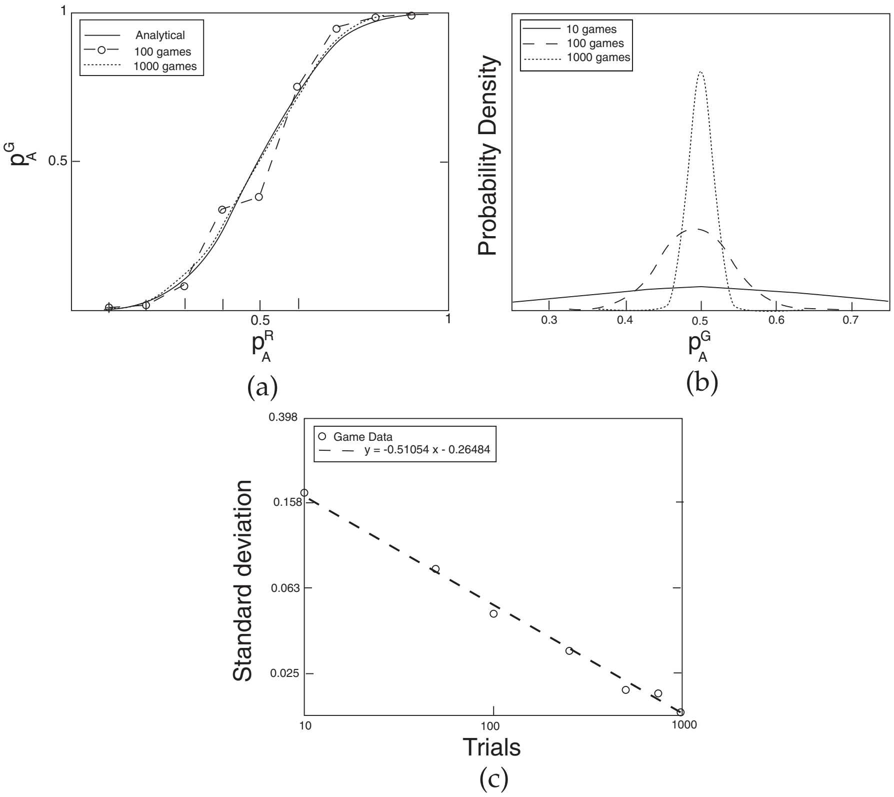
omit $j-1$ term if $i = 7$, $j \le 6$;
omit $i-1$ term if $j = 7$, $i \le 6$.

These must be supplemented with the initial conditions

$$p_A^T(0,0) = 1, \quad p_A^T(i,j) = 0 \tag{1.10}$$

if $i < 0$ or $j < 0$.

We then calculate p_A^T by using the solution of (1.8)–(1.10) in (1.7). This allows us to calculate p_A^S by using the recursion equations after (1.2) along with (1.3) and (1.4), with the result for p_A^T, in (1.2). More details along with all the solutions of the recursion formulas can be found in Part I. As long as p_A^R and p_B^R are held fixed, the approach is equivalent to a Markov chain model [9]. The analytical curves which are the results from this theory are shown and discussed in the next section. The main point to make is that the formulas for winning a tiebreaker, set, and match for each player are functions of *both* p_A^R and p_B^R, in contrast to the formula (1.1) for winning a game. This makes it much more difficult to gather large quantities of data for comparison purposes.



(a)

(b)

(c)

Fig. 3 *Convergence of the Monte Carlo simulation (*100 *trials and* 1000 *trials) to the analytical curves (solid).* (a) p_A^G *vs.* p_A^R*;* (b) *Gaussian distributions showing convergence to the mean for* 10, 100, *and* 1000 *trials;* (c) *standard deviation as a function of the number of games showing power law convergence* $\sigma \sim \alpha n^{-\beta}$ *with convergence rate* $\beta \sim 0.511$.

1.2. Description of the Monte Carlo Approach. The starting point for a Monte Carlo simulation of tennis is a random number generator which is capable of generating values for p_A^R and p_B^R sampled from a uniform distribution on the interval $[0,1]$. When player A is serving, for each point we sample a value on the unit interval, and if the value lies in the range $[0,p_A^R]$, player A wins the point, otherwise player B wins the point. Similarly, when player B is serving, for each point we sample a value on the unit interval, and if the value lies in the range $[0,p_B^R]$, player B wins the point, otherwise player A wins the point. The point-by-point simulation proceeds in this way according to the scoring rules of tennis (see Part I), and statistics are gathered to show the number of games, sets, and matches won by each. For our purposes, the pseudorandom number generation algorithm RAND in MATLAB is suitable and the statistics generated from a sequence of trials is discussed below. The average computational time per match is roughly 3 ms.

2. Convergence to Analytical Theory. Figure 3 shows the convergence results for games as a function of points. We show the analytical curve from Part I, together with the statistics based on 100 realizations and 1000 realizations of games (Figure 3(a)). Figure 3(b) shows the Gaussian convergence for one of the data points ($p_A^R =$

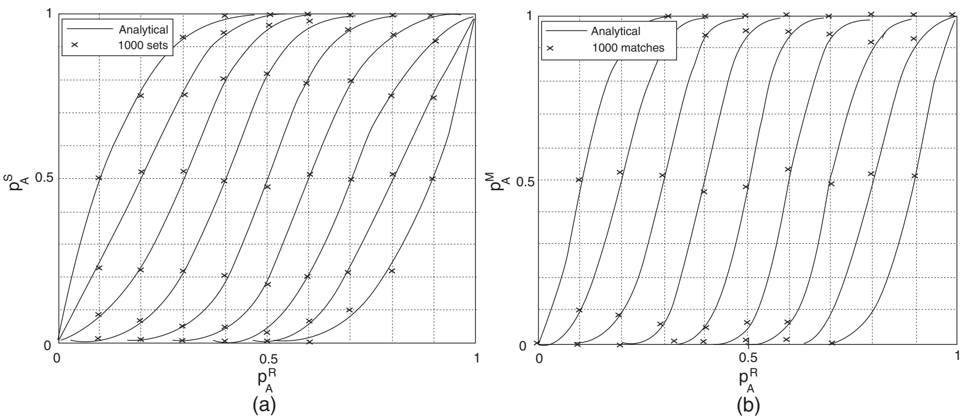


Fig. 4 (a) p_A^S *vs.* p_A^R *showing convergence of the Monte Carlo simulation (*1000 *trials) to the analytical curves (solid) against a spectrum of servers* $0.0 < p_B^R < 1.0$*;* (b) p_A^M *vs.* p_A^R *showing convergence of the Monte Carlo simulation (*1000 *trials) to the analytical curves (solid) against a spectrum of servers* $0.0 < p_B^R < 1.0$.

0.5) which converges most slowly to the correct analytical value ($p_A^G = 0.5$) for 10, 100, and 1000 trials. Figure 3(c) shows the standard deviation σ as a function of the number of games plotted on a log-log scale. The data shows a characteristic power law convergence $\sigma \sim \alpha n^{-\beta}$ with power law exponent $\beta \sim 0.511$. The convergence for sets and matches is similarly rapid and of the same power law form, with exponents $\beta \sim 0.611$ and $\beta \sim 0.476$, respectively. In all cases we found that after 1000 realizations, convergence was sufficiently close to the analytical curves and was uniform throughout the entire range of values of p_A^R. Thus, in practice, this relatively small number of realizations was used.

Figure 4(a) shows the results of p_A^S as a function of p_A^R for the entire spectrum of opponents; hence, $p_B^R = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$. Again, for 1000 realizations, the convergence to the analytical curves is uniform throughout the entire range. The same is true of Figure 4(b), which shows results for p_A^M (3 out of 5 set format) as a function of p_A^R against a wide spectrum of opponents. The results for the 2 out of 3 set format are similar. Our main conclusion from these test runs is that 1000 trials are sufficient for most purposes to ensure that the statistics are accurate throughout the full range of both parameters p_A^R and p_B^R.

3. Non-i.i.d. Models.

To test the robustness of the analytical model based on the i.i.d. assumption, we can perturb the values of p_A^R and p_B^R to see what effect this has on the shapes of the curves. We do this in the context of modeling non-i.i.d. effects.

3.1. Importance.

Not all points in a tennis match are equally important to determining its outcome. Although this is well known to professional players, there is not uniform agreement on which points are the most important. Some think that the first point of a game is the most important as it is crucial to get off to a good start, while probably the points most frequently cited as being pivotal are 15-30 and 30-15. What is agreed upon is that great players adjust their efforts according to which points, games, and sets they feel are the most crucial toward winning a match. In fact, one of the qualities that is frequently cited as a sign of a great champion is his or her ability to focus on key points and to be able to raise the level of play accordingly. The most recent example of a player who seemed to have this ability is Pete Sampras, but Bjorn Borg and Chris Evert are also frequently singled out.

If, indeed, players changed their level of effort according to the point, game, or set, then not all points, games, or sets would be identical and the i.i.d. assumption on which the theory in Part I is based would need to be modified. To carry out this modification, we describe a method for quantifying the importance of a point, game, or set based on an original formulation of Morris [14] and examined further in Pollard [19]. Morris defined the *importance* of a point for winning the game as the difference between two conditional probabilities: the probability that the server wins the game given that he wins the point, minus the probability that he wins the game given that he loses the point. If I_{ij}^{p} denotes the importance of point i for the server and j for the receiver for winning the game, then

$$I_{ij}^{P} = P_{i+1,j}^{G} - P_{i,j+1}^{G}, \tag{3.1}$$

where we define P_{ij}^{G} as the probability that the server will win the game given that the score is i points for the server and j points for the receiver. In a similar way, we can define the importance of a given game toward winning a set as

$$I_{ij}^{G} = P_{i+1,j}^{S} - P_{i,j+1}^{S}, \tag{3.2}$$

where P_{ij}^{S} denotes the probability that the first server will win the set given that the score is i games for the first server and j games for the first receiver. The importance of each point toward winning a tiebreaker, I_{ij}^{T}, is defined as

$$I_{ij}^{T} = P_{i+1,j}^{T} - P_{i,j+1}^{T}, \tag{3.3}$$

where we define P_{ij}^{T} as the probability that the first server will win the tiebreaker given that the score is i points for the first server and j points for the first receiver. Finally, the importance of each set toward winning a match is defined as

$$I_{ij}^{S} = P_{i+1,j}^{M} - P_{i,j+1}^{M}, \tag{3.4}$$

where P_{ij}^{M} denotes the probability that the first server will win the match given that the score is i sets for the first server and j sets for the first receiver. We can then calculate the importance of each point, game, and set in terms of the variables p_{A}^{R} and p_{B}^{R}.

To obtain the terms for the importance of each point, we solve hierarchically the system

$$P_{ij}^{G} = p_{A}^{R} P_{i+1j}^{G} + q_{A}^{R} P_{ij+1}^{G}, \quad i,j = 0,1,2, \tag{3.5}$$

and when $i = 3$ or $j = 3$ we use

$$P_{31}^{G} = p_{A}^{R} + q_{A}^{R} P_{32}^{G}, \tag{3.6}$$

$$P_{13}^{G} = p_{A}^{R} P_{23}^{G}, \tag{3.7}$$

$$p_{30}^{G} = p_{A}^{R} + q_{A}^{R} P_{31}^{G}, \tag{3.8}$$

$$p_{03}^{G} = p_{A}^{R} P_{13}^{G}. \tag{3.9}$$

Since the probability of winning a game when the score is 0 to 0 is simply p_{A}^{G}, we have as an initial condition

$$P_{00}^{G} \equiv p_{A}^{G} = (p_{A}^{R})^{4}[1 + 4q_{A}^{R} + 10(q_{A}^{R})^{2}] + 20(p_{A}^{R}q_{A}^{R})^{3}(p_{A}^{R})^{2}[1 - 2p_{A}^{R}q_{A}^{R}]^{-1}. \tag{3.10}$$

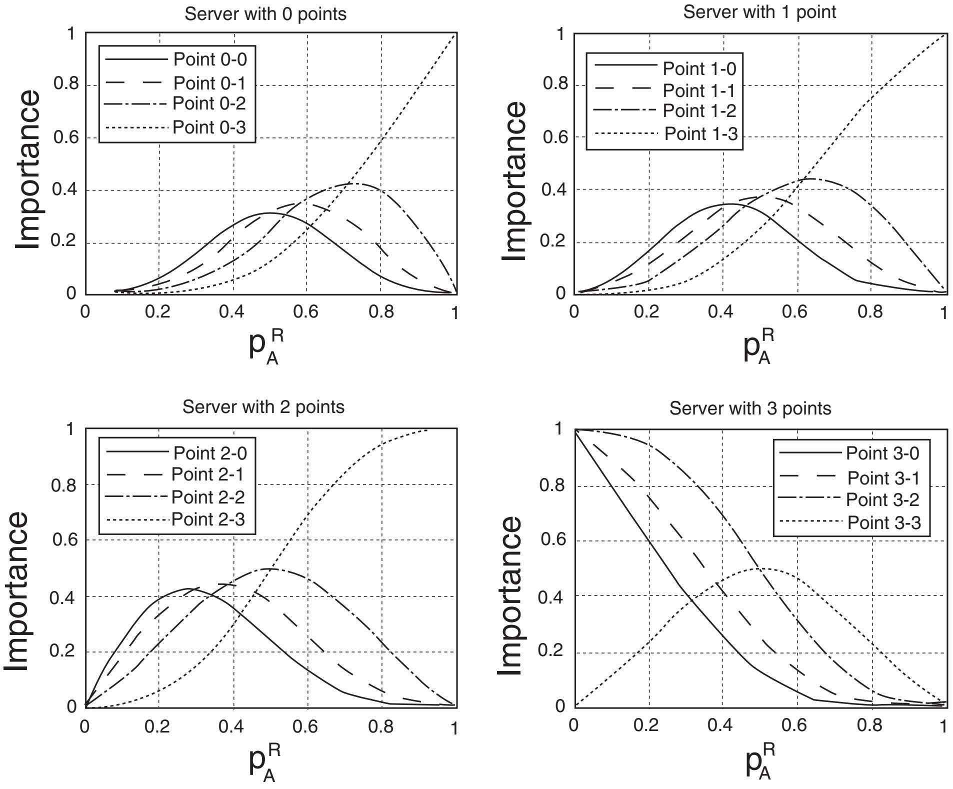


Fig. 5 *Importance of each point in terms of winning a game as a function of p_A^R.*

We show the results for the importance of each point in Figure 5. Each figure shows the importance curves as a function of p_A^R with the server having 0, 1, 2, or 3 points, respectively. The curves show that the importance of the point depends on the value of p_A^R and how many points the server currently has. Figure 6 shows the point with maximal and minimal importance as a function of p_A^R. In the region $0 \leq p_A^R \leq 0.5$, the most important point is 40-30 and the least important point is 0-40, whereas in the region $0.5 \leq p_A^R \leq 1.0$, which is the typical range for professional players, the most important point is 30-40, while the least important point is 40-0.

The importance of each point in a tiebreaker, game, and set can be obtained by solving similar hierarchical systems of equations, taking care to keep track of the alternating service games between players A and B throughout the set and the serving order in a tiebreaker. For example, if playing against Pete Sampras, the importance of each point in a tiebreaker and the importance of each game toward winning a set as a function of p_A^R are shown in Figures 7 and 8, respectively. Again, the detailed values of the importance of the various points in a tiebreaker and games in a set depend on the values of both p_A^R and p_B^R.

We can now use this information to test the effects of a non-i.i.d. model based on the following idea. Suppose player A adjusts his or her level of play according to the importance of each point. In the simplest case, we adjust player A's nominal value of p_A^R by 20%, increasing the value on the most important point of the game (40-30

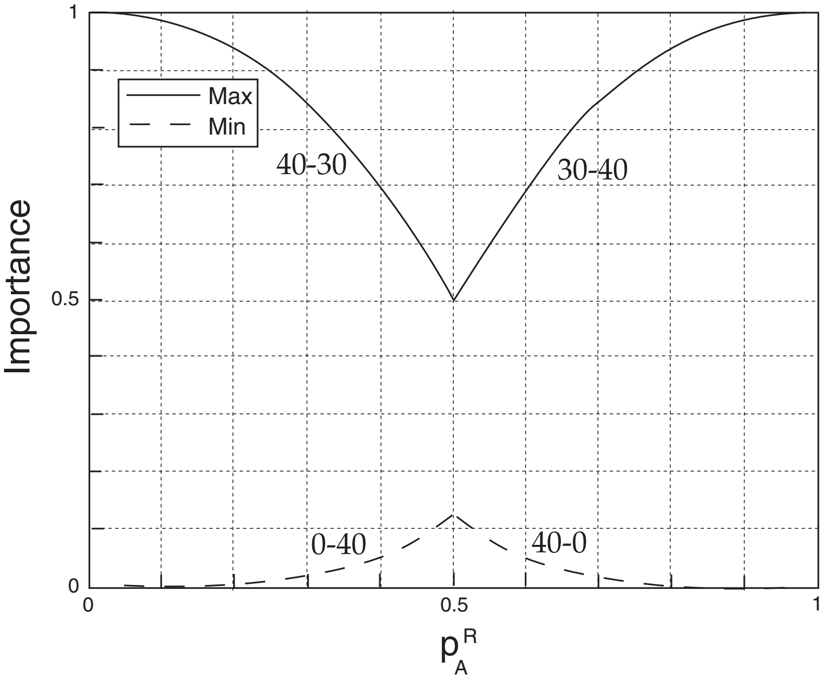


Fig. 6 *Curves of maximal (solid) and minimal (dashed) importance as a function of* p_A^R.

or 30-40), and decreasing the value by the same amount on the least important point (40-0 or 0-40). Note that the least important point in a game can occur only once, whereas the most important point can occur arbitrarily many times (ad-in, which means the server wins the next point after deuce, is equivalent to 40-30 and ad-out, which means the server loses the next point after deuce, is equivalent to 30-40). The effect of this model is shown in Figure 9 for the values $p_B^R = 0.2, 0.4, 0.6, 0.8$ and for the full range of values of p_A^R. The overall effect is that the curves are shifted slightly up from what they would have been in the pure i.i.d. theory, i.e., player A's probability of winning is increased in this model. The reason for this is that the adjustment *up* on the most important point occurs more frequently than the adjustment *down* on the least important point; hence the effective value for p_A^R is slightly higher than that in the baseline i.i.d. case.

3.2. The Hot-Hand Effect: Does Success Breed Success? To model the hot-hand effect, we perturb each player's value of p_A^R or p_B^R on the one point immediately following each point that they win. Figure 10 shows the result of perturbations with 20% amplitude taken from a uniform distribution evenly distributed around the analytical curve. The figure shows relatively rapid convergence to the i.i.d. curves for 10, 100, and 1000 trials. After 1000 trials, the convergence to the analytical curves is uniform throughout the range. This is somewhat surprising given the size of the perturbations and their non-i.i.d. nature; however, because they are taken from a uniform distribution which is symmetric about the analytical curves, the increased kicks and decreased kicks from the nominal analytical values of p_A^R and p_B^R effectively cancel each other out after a sufficiently large number of trials. Contrast this with

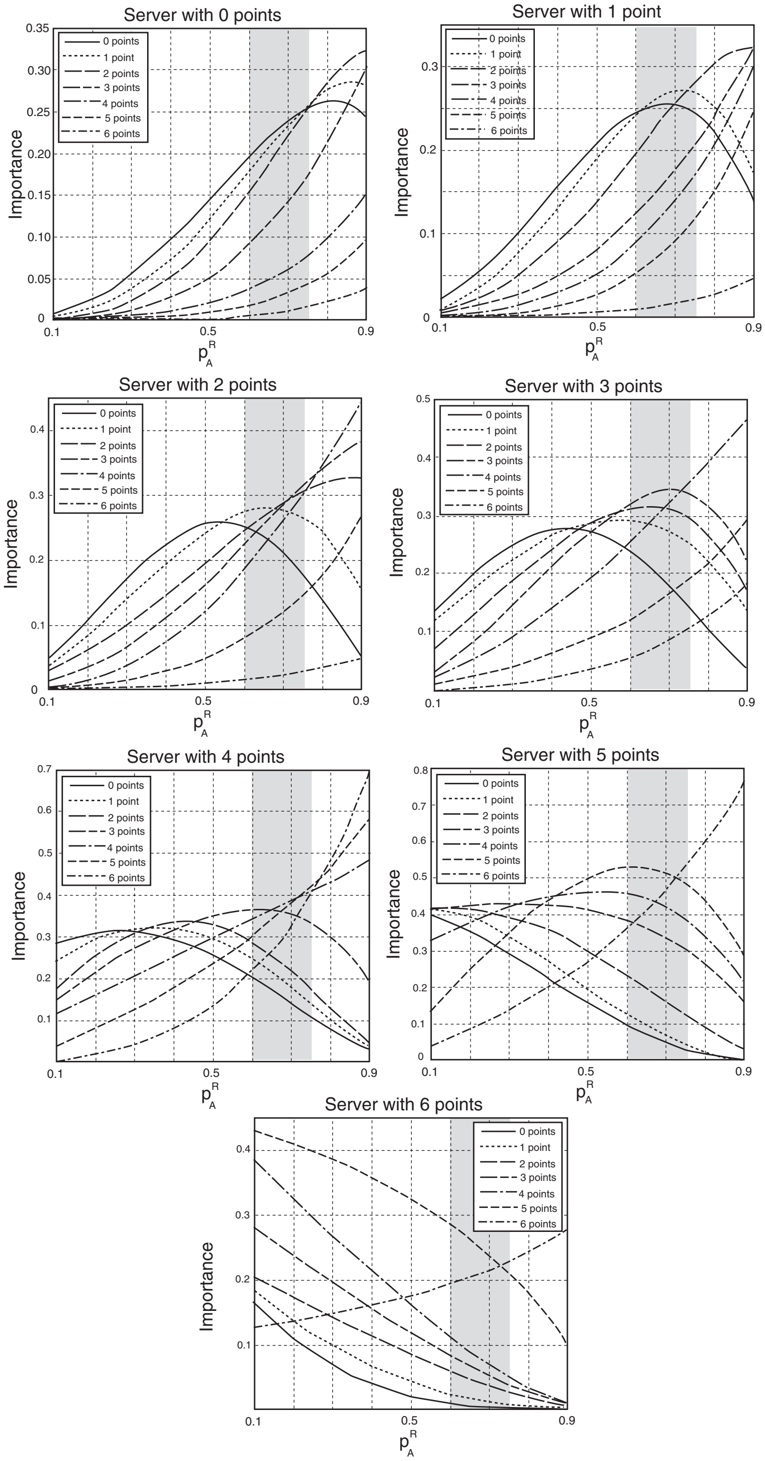


Fig. 7 *Importance of points in a tiebreaker against Pete Sampras* ($p_B^R = 0.73$)*. Gray zone marks the typical professional range* $0.6 < p_A^R < 0.75$*.*

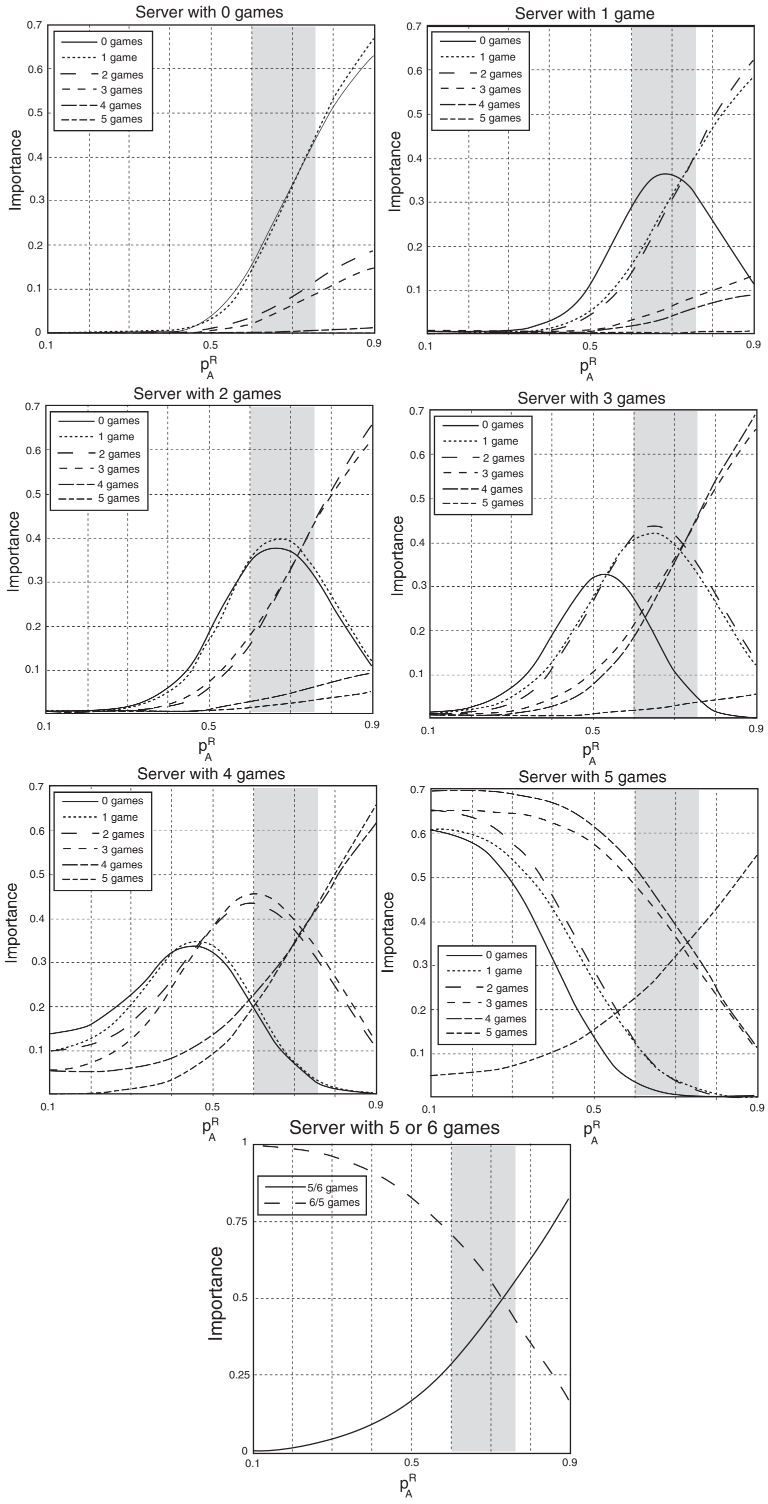


Fig. 8 *Importance of games toward winning a set against Pete Sampras* ($p_B^R = 0.73$). *Gray zone marks the typical professional range* $0.6 < p_A^R < 0.75$.

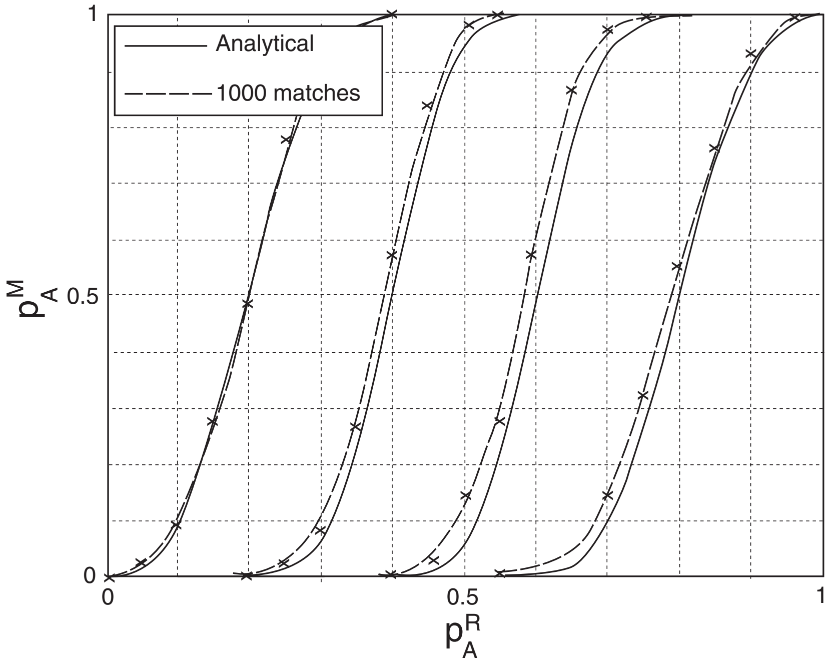


Fig. 9 *Non-i.i.d. effects based on adjusting play according to the importance of points for a range of values of* p_B^R.

Figure 12, which shows the results from hot-hand perturbations that are are still large (20% size amplitudes) but are not symmetric about the analytical curve. Here, we *increase* the nominal value of p_A^R or p_B^R, depending on who is serving, on each point after a point is won. The result, after 1000 trials, does not converge to the i.i.d. curves but shows a systematic shift upwards. Hence, the server's probability of winning a game is increased over what it would be from the pure i.i.d. theory.

3.3. Back-to-the-Wall Effect: Does Failure Breed Success?

To model the back-to-the-wall effect, we perturb each player's value of p_A^R or p_B^R by a fixed percentage on one point immediately following each point that they lose. Figure 11 shows the result of perturbations with 20% amplitude taken from a uniform distribution evenly distributed around the analytical curve. The figure again shows relatively rapid convergence to the i.i.d. curves for 10, 100, and 1000 trials. After 1000 trials, the convergence to the analytical curves is uniform throughout the range. Figure 12 shows the results from back-to-the-wall perturbations that are are still large (20% size amplitudes) but are not symmetric about the analytical curve. Here, we increase the nominal value of p_A^R or p_B^R, depending on who is serving, on each point after a point is lost. Again, the result, after 1000 trials, does not converge to the i.i.d. curves but shows a systematic shift upwards. Hence, as in the hot-hand perturbations, the server's probability of winning a game is increased over what it would be from the pure i.i.d. theory.

4. Arrow's Impossibility Theorem and Probabilistic Ranking Systems.

Sports ranking systems take many different forms, as reviewed in Stefani [22], but, roughly speaking, they can be grouped into two distinct categories. One type attempts to

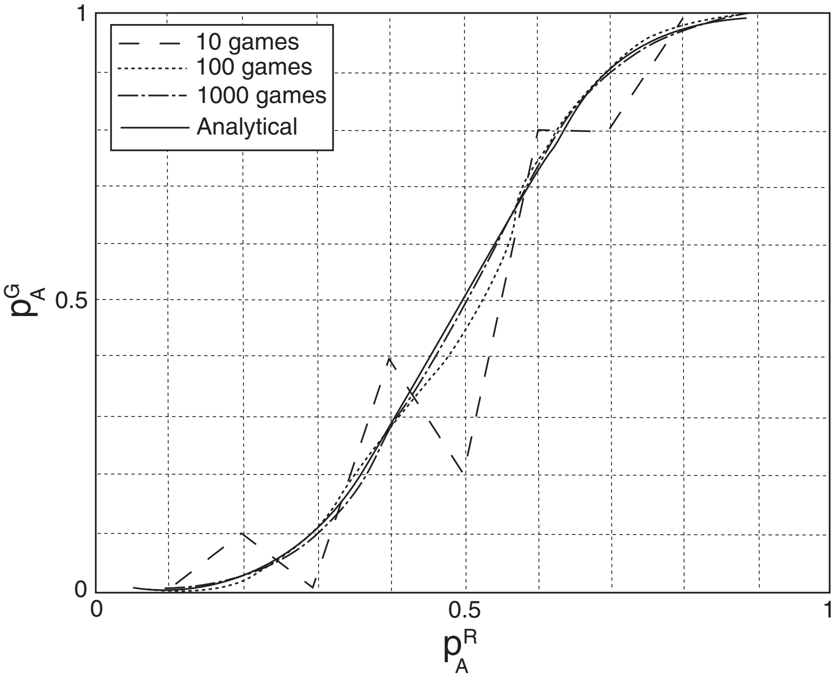


Fig. 10 *Random white noise hot-hand perturbations with 20% amplitude.*

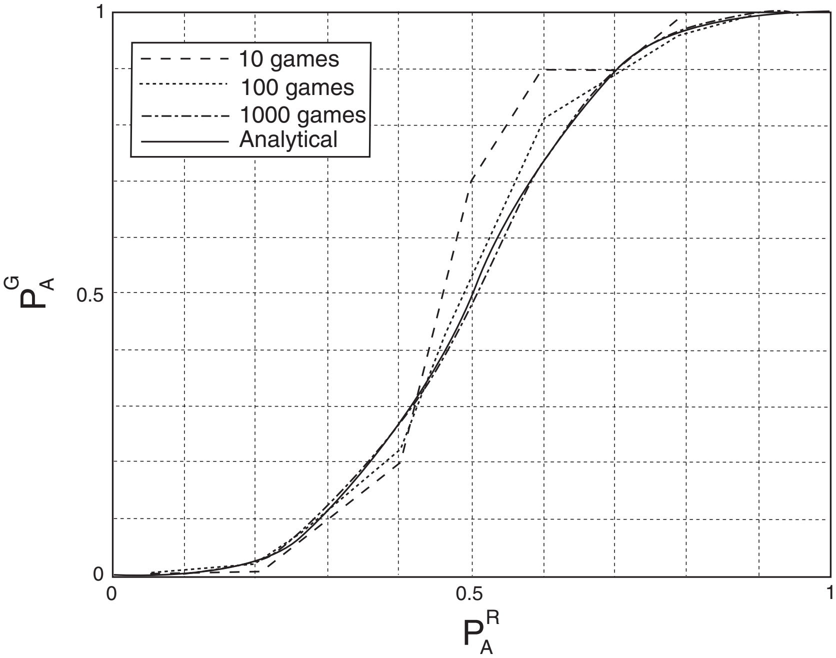


Fig. 11 *Random white noise back-to-the-wall perturbations with 20% amplitude.*

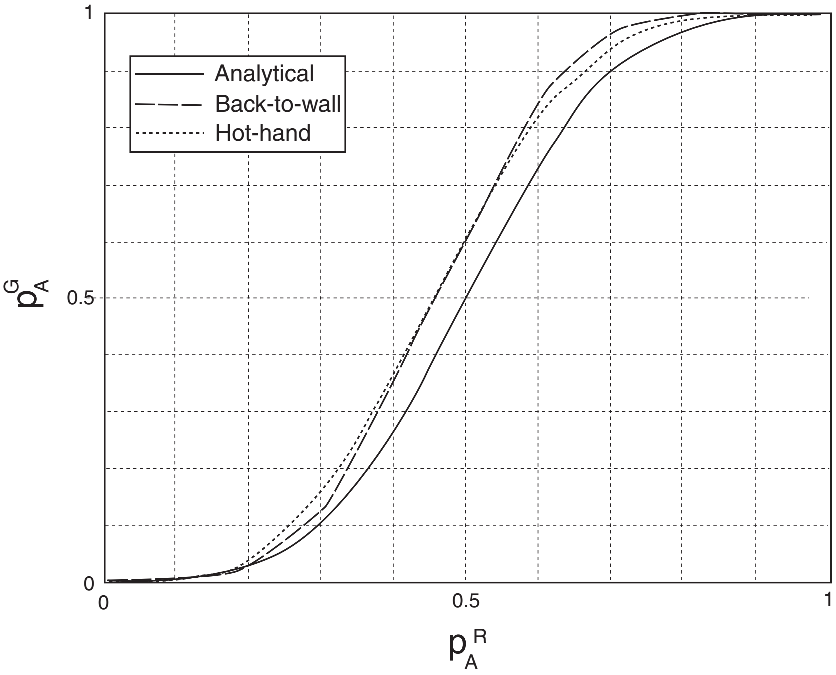


Fig. 12 *Curves showing the effect of hot-hand perturbations and back-to-the-wall perturbations with* 20% *amplitude.*

reward teams after a predetermined number of games are played, by, for example, crowning them as national champions or year-end champions—we call these types of ranking systems "outcome based." The philosophy behind these methods is that there *should* be a player or team who "deserves" to be recognized as "the best," and if only the correct method were found, such a team or player could be unambiguously chosen. Examples of systems of this type are the ones currently used in college football (see [4] for a lucid description of the BCS system) and in men's and women's professional tennis. They typically rely on matrix methods of the type surveyed nicely in Keener [8] and used by Colley [6] for ranking football teams.

With regard to this use of ranking systems, it is an underappreciated fact that Arrow's impossibility theorem [2] for aggregating individual preferences into social preferences holds. The theorem states that *there is no rule*, majority voting or otherwise, for establishing social preferences from arbitrary individual preferences (the monkey ranking system notwithstanding [4, 5]). In other words, under certain conditions of rationality and equality, it is *impossible* to guarantee that a ranking of societal preferences will correspond to rankings of individual preferences when more than two individuals and alternative choices are involved.

As a simple and concrete example of how this translates into the ability of outcome-based methods to unambiguously determine rankings, consider the problem of trying to decide upon the year-end men's tennis champion for the 2002 season in which Pete Sampras won the U.S. Open, Andre Agassi won the Australian Open, and Lleyton Hewitt won Wimbledon. Suppose we do this by majority voting among three judges between each pair of players. Let one of the judges be an American who might naturally favor the U.S. Open champion, one an Australian who favors the Aus-

Table 1 *Fictional year-end voting by three judges in the* 2002 *men's professional tennis tour. Each player won one of the grand slam events.*

Judge	Sampras	Agassi	Hewitt
American	1	2	3
British	2	3	1
Australian	3	1	2

tralian Open champion, and the third British, who favors the Wimbledon champion. A set of preferences is said to be rational (or transitive) if when player A is ranked higher than player B and B is ranked higher than C, then A is ranked higher than C. Certainly this is a desirable property for any ranking system and any system that produces an outcome without this property is likely to be viewed as unfair. Consider Table 1, which shows rankings chosen by the three judges. Suppose the American judge gives Sampras his number one rating, the British judge gives Hewitt his top rating, and the Australian judge gives Agassi his number one rating. The number two and three rankings for each of the judges are also shown in the table. If we compile the final rankings by majority vote, in a choice between Sampras and Agassi, since two out of the three judges voted Sampras as the higher-ranked player, he would be ranked above Agassi. In a choice between Agassi and Hewitt, Agassi would be ranked ahead of Hewitt since two out of the three judges voted this way. Now, since Sampras was chosen ahead of Agassi, and Agassi ahead of Hewitt, logic would tell us to rank Sampras ahead of Hewitt, i.e., transitivity should hold. But consider the outcome of the voting among the three judges in trying to decide between Sampras and Hewitt—two out of the three of the judges ranked Hewitt higher than Sampras. The outcome is irrational as transitivity does not hold.

Despite the fact that Kenneth Arrow was awarded the Nobel Prize in Economics in 1972 for his work, there continues to be a widespread belief that if the right method were discovered, there would be a "correct" way to crown a national champion in college football or a year-end champion in tennis which would eliminate irrational outcomes and settle all arguments, leaving everyone satisfied. But in fact Arrow proved that under certain reasonable assumptions, there is *no* method for constructing social preferences (rankings) from arbitrary individual ones (votes). Such outcome-based methods based on voting, as the one used to crown the NCAA national football champion, very often produce logical inconsistencies that are the basis for arguments that cannot be settled rationally. The amount of energy and effort spent on arguing over rankings of all types (particularly in college football, where few games are played compared to the total number of teams involved, but also regarding the notorious U.S. News and World Report Annual Ranking of Colleges) is an indication of the pervasiveness of Arrow's theorem.

A second, and very different, kind of system attempts to use rankings for the purpose of predicting outcomes; hence we call these "predictive methods." In a sense, they are inherently probabilistic. This type of system is used, for example, by the amateur and professional chess tour where rankings are designed so that they can be directly translated into predictions of outcomes in head-to-head competitions (see [22]). In tennis, it is widely recognized that this is not possible to do with the current ranking systems. In fact, two of the four grand slam tournaments (the French Open and Wimbledon) do not use the world rankings when it comes to seeding the players in the draw before the tournament begins. One then has to wonder what sort of ad hoc system they adopt when choosing the seed, as this choice can have very direct and important consequences on how far players advance in the draw, which, in turn, affects their future seedings, rankings, and earnings.

For example, the seedings of the men's 2005 Wimbledon draw created a controversy when the number-two-ranked player in the world, Lleyton Hewitt, was seeded third, causing him to face the top-ranked player, Roger Federer, in the semifinal round instead of the finals. His loss to Federer in the semifinals instead of the finals cost Hewitt several hundreds of thousands of dollars, since players are paid according to how far they advance in the tournament. With regard to world rankings, on two separate recent occasions in women's professional tennis, the top ranking went to players who had never won a grand slam championship.

A way to soften the consequences of Arrow's theorem is to represent the alternatives as elements in a spectrum of possibilities, i.e., to use probability-based ranking systems. Then, if the preferences of the individual exhibit *single-peakedness*, the societal preferences can be constructed unambiguously. The Monte Carlo methods described in this paper have the potential to provide such *distributional* information; hence they represent an important step in the direction of designing a probabilistic ranking system for tennis. The idea is to run thousands of simulated tournaments with players randomly ordered in fictitious draws before the tournament begins and then use the accumulated statistical winning distributions as the basis for seeding the actual tournament before it is played. The player who is most likely to win the tournament based on the simulations would be the number one seed in the real draw, the player with the second highest winning percentage would be seeded second, and so on. The input for each player, at the very least, would be his or her accumulated ratio of points won on serve to points served, together with higher-order fluctuations (thus allowing for non-i.i.d. effects), collected over the most recent relevant tournaments (i.e., clay court tournaments would be used to seed the French Open, whereas grass court tournaments would be used to seed Wimbledon). A more elaborate vector of input parameters for each player (for example, their ability to win points on service returns) could also be used. Moreover, we believe the computational power is available to run similar simulations in other sports, such as NCAA College Basketball and Football. Moving toward systems that are probabilistic and predictive would finesse the inherent inconsistencies guaranteed by Arrow's impossibility theorem and would better reflect the reality that a victory of a lower-ranked player over a higher-ranked one in a *single match* is not necessarily an inconsistent outcome.

5. Discussion.

We finish by showing one use of these Monte Carlo simulations for the purpose of predicting tournament outcomes based on data gathered throughout a tournament or in previous tournaments. Figures 13 and 14 show the 2002 U.S. Open men's and women's draws from the semifinals onward, with the values of p_A^R and p_B^R collected for each of the four players over the previous rounds of the tournament. These values are listed under each player's name as $p_i^R(n)$, where $i = 1, 2, 3, 4$ indicates the player and n indicates the number of previous rounds over which the data was collected. Using the values $p_i^R(5)$ for each of the players, we then run 1000 simulated matches for each of the next two rounds and gather statistics giving the probabilities for each player to advance to the next round (P_{ij}) and for each to become the ultimate tournament champion (superscript "TC"). In each case, the error bars with one standard deviation are shown as well. The comparison with the numbers from the analytical theory from Part I are not shown, but in each case the analytical predictions are within one standard deviation of the simulated values.

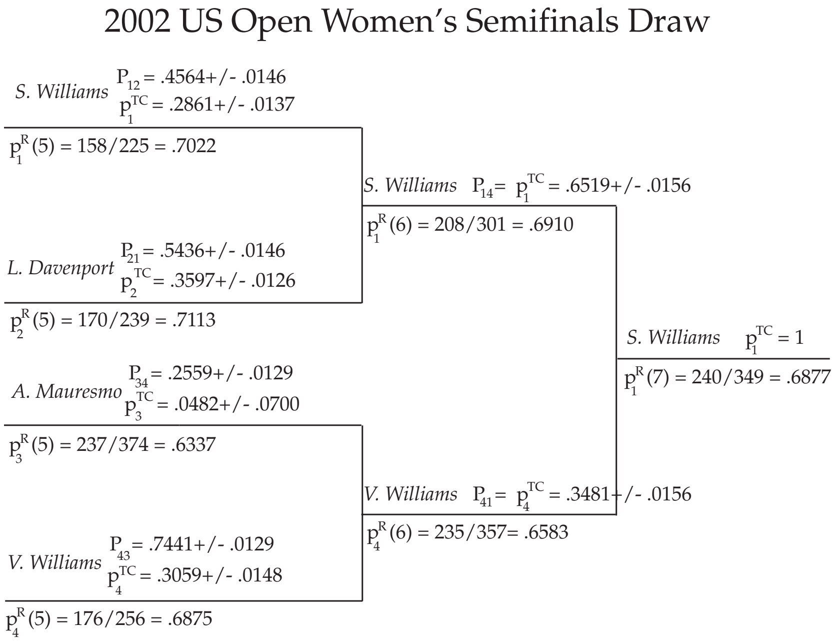

2002 US Open Women's Semifinals Draw
S. Williams
P_{12} = .4564+/- .0146
p_1^{TC} = .2861+/- .0137
$p_1^R(5)$ = 158/225 = .7022
L. Davenport
P_{21} = .5436+/- .0146
p_2^{TC} = .3597+/- .0126
$p_2^R(5)$ = 170/239 = .7113
A. Mauresmo
P_{34} = .2559+/- .0129
p_3^{TC} = .0482+/- .0700
$p_3^R(5)$ = 237/374 = .6337
V. Williams
P_{43} = .7441+/- .0129
p_4^{TC} = .3059+/- .0148
$p_4^R(5)$ = 176/256 = .6875
S. Williams P_{14} = p_1^{TC} = .6519+/- .0156
$p_1^R(6)$ = 208/301 = .6910
V. Williams P_{41} = p_4^{TC} = .3481+/- .0156
$p_4^R(6)$ = 235/357= .6583
S. Williams p_1^{TC} = 1
$p_1^R(7)$ = 240/349 = .6877

Fig. 13 *Tournament predictions based on Monte Carlo simulations of the* 2002 *U.S. Open women's draw using data through the quarterfinal round. Note that the eventual champion, S. Williams, had only a* 28.61% *chance of winning the tournament based on data through the quarterfinal round, while L. Davenport and V. Williams had higher percentages. However, after the semifinal round, her chances of winning the tournament increased to* 65.19%*, making her the clear favorite.*

We believe this gives a glimpse into the power of the Monte Carlo code and its potential usefulness and flexibility both in doing full tournament simulations while holding the values of p_A^R and p_B^R for each player fixed, and also for doing non-i.i.d. simulations while varying these values in some prescribed way. One of the main conclusions of our work, however, is that varying these values in ways that might be considered reasonable from the point of view of modeling non-i.i.d. effects such as hot-hand, back-to-the-wall, or random fluctuations does not dramatically alter the probabilities predicted from a pure i.i.d. theory (see discussions of this effect in [21]). While this result may be somewhat surprising, it is consistent with many of the previously cited results that indicate the difficulties in detecting non-i.i.d. effects in data sets, not only in tennis but in other sports as well. It highlights the *unreasonable effectiveness* of the i.i.d. assumption in certain situations even when it is suspected that non-i.i.d. effects are present. While the i.i.d. assumption may not be perfect, it is important to remember that even if one knows that non-i.i.d. effects are present, the choice of a particular non-i.i.d. model can introduce further sources of error.

The Monte Carlo program for running tennis simulations is a working, evolving code written in the MATLAB programming environment that can be used to shed light on many questions one could think of, and we welcome such questions from interested readers. Our code is currently being used to simulate a grand slam season in tennis and to develop probabilistic tennis ranking systems for both men and women.

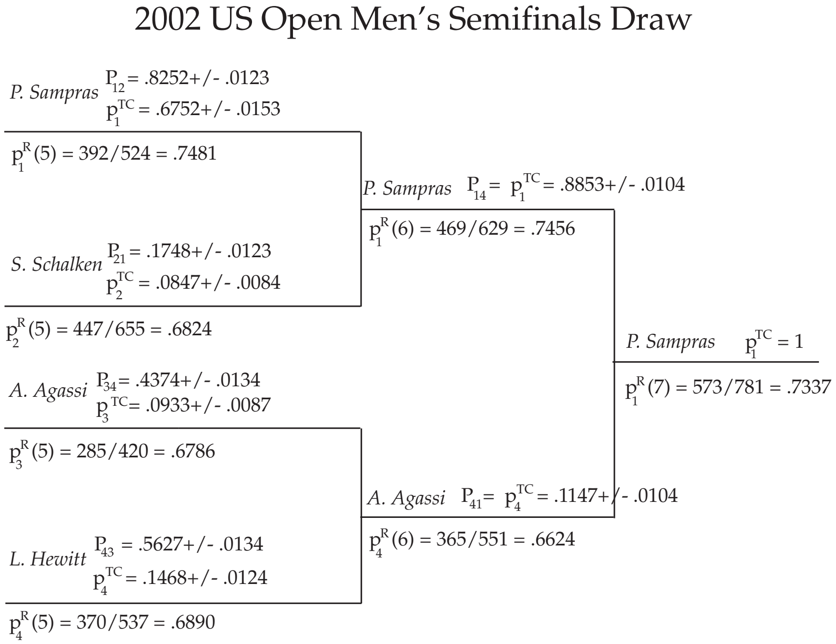

2002 US Open Men's Semifinals Draw
P. Sampras P_{12}= .8252+/- .0123
p_1^{TC} = .6752+/- .0153
p_1^R(5) = 392/524 = .7481
S. Schalken P_{21} = .1748+/- .0123
p_2^{TC} = .0847+/- .0084
p_2^R(5) = 447/655 = .6824
A. Agassi P_{34}= .4374+/- .0134
p_3^{TC}= .0933+/- .0087
p_3^R(5) = 285/420 = .6786
L. Hewitt P_{43} = .5627+/- .0134
p_4^{TC}= .1468+/- .0124
p_4^R(5) = 370/537 = .6890
P. Sampras P_{14}= p_1^{TC}= .8853+/- .0104
p_1^R(6) = 469/629 = .7456
A. Agassi P_{41}= p_4^{TC}= .1147+/- .0104
p_4^R(6) = 365/551 = .6624
P. Sampras p_1^{TC} = 1
p_1^R(7) = 573/781 = .7337

Fig. 14 *Tournament predictions based on Monte Carlo simulations of the* 2002 *U.S. Open men's draw using data through the quarterfinal round. Based on this data, the tournament champion, P. Sampras, was the clear favorite, with a* 67.52% *chance of winning the tournament after the quarterfinal round. His probability of winning the tournament then increased to* 88.53% *after his semifinal performance, making him the prohibitive favorite over A. Agassi, whose chance of winning the finals was only* 11.47%.

Acknowledgments. The first author would like to thank J. B. Keller, P. J. Mucha, A. L. Ruina, and R. T. Stefani for stimulating and useful conversations throughout the course of this work.

REFERENCES

[1] S.C. Albright, *A statistical analysis of hitting streaks in baseball*, J. Amer. Statist. Assoc., 88 (1993), pp. 1175–1183.

[2] K. Arrow, *Social Choice and Individual Values*, 2nd ed., Yale University Press, New Haven, CT, 1970.

[3] W.H. Carter and S.L. Crews, *An analysis of the game of tennis*, Amer. Statist., 28 (1974), pp. 130–134.

[4] T. Callahan, P.J. Mucha, and M.A. Porter, *The Bowl Championship Series: A Mathematical Review*, Notices Amer. Math. Soc., 51 (2004), pp. 887–893.

[5] T. Callahan, P.J. Mucha, and M.A. Porter, *Random walker ranking for Division I-A football*, Amer. Math. Monthly, to appear.

[6] W.N. Colley, *Colley's Bias Free College Football Ranking Method: The Colley Matrix Explained*, preprint.

[7] D. Jackson and K. Mosurski, *Heavy defeats in tennis: Psychological momentum or random effects*, Chance, 10 (1997), pp. 27–34.

[8] J.P. Keener, *The Perron–Frobenius theorem and the ranking of football teams*, SIAM Rev., 35 (1993), pp. 80–93.

[9] J.G. Kemeny and L. Snell, *Finite Markov Chains*, Springer-Verlag, New York, 2004.

[10] F.J.G.M. Klaassen and J.R. Magnus, *Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model*, J. Amer. Statist. Assoc., 96 (2001), pp. 500–509.

[11] J.R. Magnus and F.J.G.M. Klaassen, *On the advantage of serving first in a tennis set: Four years at Wimbledon*, The Statistician, 48 (1999), pp. 247–256.

[12] J.R. Magnus and F.J.G.M. Klaassen, *The effect of new balls in tennis: Four years at Wimbledon*, The Statistician, 48 (1999), pp. 239–246.

[13] J.R. Magnus and F.J.G.M. Klaassen, *The final set in a tennis match: Four years at Wimbledon*, J. Appl. Statist., 26 (1999), pp. 461–468.

[14] C. Morris, *The most important points in tennis*, in Optimal Strategies in Sport, S.P. Ladany and R.E. Machol, eds., North-Holland, Amsterdam, The Netherlands, 1977, pp. 131–140.

[15] P.K. Newton and J.B. Keller, *The probability of winning at tennis* I. *Theory and data*, Stud. Appl. Math., 114 (2005), pp. 241–269.

[16] P.K. Newton and G.H. Pollard, *Service neutral scoring strategies for tennis*, in Proceedings of the Seventh Australasian Conference on Mathematics and Computers in Sport, Massey University, Palmerston North, New Zealand, 2004, pp. 221–225.

[17] G.H. Pollard, *An analysis of classical and tie-breaker tennis*, Austral. J. Statist., 25 (1983), pp. 496–505.

[18] G.H. Pollard, *The effect of a variation to the assumption that the probability of winning a point in tennis is constant*, in Proceedings of the Sixth Australian Conference on Mathematics and Computers in Sport, G. Cohen and T. Langtry, eds., University of Technology, Sydney, 2002, pp. 221–226.

[19] G.H. Pollard, *Can a player increase the probability of winning a point when it is more important?*, in Proceedings of the Seventh Australasian Conference on Mathematics and Computers in Sport, Massey University, Palmerston North, New Zealand, 2004, pp. 226–230.

[20] G.H. Pollard and K. Noble, *The effect of changing the assumption that the probability of winning a point in a tennis match is constant*, in Tennis Science and Technology 2, S. Miller, ed., International Tennis Federation, London, 2003, pp. 341–346.

[21] G.H. Pollard and K. Noble, *The effect of having correlated point outcomes in tennis*, in Proceedings of the Seventh Australasian Conference on Mathematics and Computers in Sport, Massey University, Palmerston North, New Zealand, 2004, pp. 230–240.

[22] R.T. Stefani, *Survey of the major world sports rating systems*, J. Appl. Statist., 24 (1997), pp. 635–646.

[23] H. Stern and C. Morris, *Comment on "A statistical analysis of hitting streaks in baseball,"* J. Amer. Statist. Assoc., 88 (1993), pp. 1189–1194.

[24] A. Tversky and T. Gilovich, *The cold facts about the "hot hand" in basketball*, Chance, 2 (1989), pp. 16–21.