International Journal of Forecasting 15 (1999) 83–91

Are sports seedings good predictors?: an evaluation

Bryan L. Boulier, H.O. Stekler*

Department of Economics, George Washington University, Washington, DC 20052, USA

Abstract

Very little attention has been given to predicting outcomes of sporting events. While studies have examined the accuracy of alternative methods of predicting the outcomes of thoroughbred horse races, some obvious predictors of the outcomes of other sporting events have not been examined. In this paper, we evaluate whether rankings (seedings) are good predictors of the actual outcomes in two sports: (1) US collegiate basketball and (2) professional tennis. In this analysis we use statistical probit regressions with the difference in rankings as the predictor of the outcome of games and/or matches. We evaluate both the ex post and ex ante predictions using base rate forecasts and Brier scores. We conclude that the rankings, by themselves, are useful predictors and that the probits improve on this performance.

Keywords: Probability forecasts; Probits; Sports forecasts

1. Introduction

Although there have been many evaluations of both weather and macroeconomic forecasts, there have been few evaluations of the predictions of the outcomes of sporting events. Exceptions are studies of thoroughbred horse races (cf. White et al., 1992) and point spreads in professional football and basketball games (Sauer et al., 1988; Camerer, 1989; Brown and Sauer, 1993). There have been no studies that evaluated procedures for using rankings (seedings) to predict the outcomes of sporting events. Here we examine the accuracy of rankings in predicting the outcomes of (1) US collegiate basketball games and (2) professional tennis matches. Our analysis of the accuracy of the seedings in predicting the outcomes of *particular* basketball games differs from the Schwertman et al. (1991) study. That paper presents alternative statistical models, based on assumptions about the information contained in the rankings, that derive the probability that any given seed will win the entire tournament.

The next section describes the sporting events that we study. We then present some descriptive results that characterize the relationship between the rankings and the relative frequencies of winning. These descriptive results yield a relationship between the relative difference in rankings and the relative frequency of winning. Using this insight we develop a statistical methodology, probit functions, for predicting the probability that a particular ranked team or player will win the event. The accuracy of the forecasts obtained from the probit statistical functions is then evaluated against a base case set of predictions.

*Corresponding author. Tel.: +1-202-994-6150; fax: +1-202-994-6147; e-mail: hstekler@gwis2.circ.gwu.edu

0169-2070/99/$ – see front matter
PII: S0169-2070(98)00067-3

2. Data (seedings)

There are many sports in which players or teams are assigned seedings or rankings. These include the National Football League playoffs leading to the Super Bowl; the Stanley Cup hockey and National Basketball Association playoffs; tennis tournaments; and various National Collegiate Athletic Association (NCAA) competitions. In many of the tournaments or playoffs, the seedings are determined by the competitors' performances during the regular season. This is partially true in the NCAA basketball tournaments and largely true in tennis. In basketball, the seeding committee uses its informed judgment as well as the prior records of the teams. In tennis the seedings are based on computer rankings derived from performances in the prior 12 months. The Wimbledon rankings are an exception to this procedure because informed judgment supplements the computer rankings in determining the seedings. In order to evaluate whether rankings are good predictors of the outcomes of sporting events, we focus on the 1985–95 competitions in the NCAA basketball and the Grand Slam tennis tournaments.

2.1. Basketball: men's and women's tournaments

Since 1985, at the end of the regular basketball season the NCAA has selected 64 college teams to participate in a tournament to select a national champion.[1] The 64 teams are divided into four groups of 16, and within each group they are (ranked) from 1 to 16. In the first round the No. 1 team plays the No. 16 in its group; No. 2 plays No. 15, etc. If the seedings were perfect predictors, all of the better-ranked teams would win in the first round, and No. 1 would then play No. 8, etc. Each group plays four rounds, with the winner from each group advancing to the Final Four Championship round. We only evaluate the predictive accuracy of the relative rankings within each group; we do not examine the Final Four Championship round.

The format of the women's college basketball tournament is currently identical to that of the men's competition, but this has not always been the case. Prior to 1994 a smaller number of teams were selected for the competition. Although there have always been four groups, in some years the top seeded teams were given byes in the first rounds. When either 32 or 64 teams were involved, all played in the first round. With 36 entrants, the No. 1 seed in each group was given a bye in the first round; with 40 teams, only teams ranked 7 through 10 played in the first set of games. Despite these differences, we will evaluate the men's and women's rankings in the same way.

2.2. Tennis: men's and women's Grand Slam events

There are four major tennis tournaments, called the Grand Slam events, which have particular significance – the Australian Open, the French Open, Wimbledon and the US Open – because almost all of the world's top-ranked players participate in these tournaments.[2] Competition begins with 128 entrants, with 16 seeded. The seeded players do not compete against each other until the fourth round, playing only unseeded entrants in the first three rounds. We, therefore, have two ways to measure the predictive accuracy of the rankings: (1) the relative frequencies that seeded players reached the fourth round and (2) the relative rankings in the fourth and later rounds as predictors of the outcome of each match. The tennis seedings are based primarily upon computer rankings derived from past performances, regardless of the surface on which previous events were played. Each player's game may be best suited to a particular surface. The four Grand Slam events are played on three different types of surfaces. Wimbledon has grass courts; the French Open is played on clay; the surfaces at the Australian and US Opens are a hard

[1] While the teams considered by the selecting committee to be among the best in the nation are in the tourney, some teams gained automatic entry to the tournament because they were the best team in their conference, e.g. the Ivy League. The conferences are composed of particular schools that play each other regularly.

[2] We analyze these Grand Slam events because their importance insures that all of the participants will compete to the best of their abilities. There has been some criticism of some players who, it is asserted, have made only token appearances at lesser events and have been eliminated in the early rounds. Using data from these events would, in our opinion, bias the results.

composite, although the Australian was played on grass until 1986. We therefore had to consider whether to pool the data from the four Grand Slam events or to analyze the events separately. The latter approach would be appropriate if the computer rankings and thus the seedings (and their predictive values) were not indicative of relative ability in one of these particular events. Another reason for considering each site separately is that Wimbledon seedings have often departed from the computer rankings. As one observer has noted, ''Wimbledon officials have always seeded at their whim...'' (Roberts, 1996, p. C5)

3. Descriptive results

The higher-ranked men's and women's basketball teams beat lower-ranked opponents in 73.5 and 77.7% of the games, respectively. The higher-ranked men's and women's tennis players beat the lower-ranked individuals in 75.8 and 81.8% of the matches, respectively. These results were compared with the binomial distribution with a 0.50 probability of picking a winner, and in all cases these ratios were statistically significant. Given that both past performance and informed judgment were used to rank teams and individuals, it is not surprising that the seedings beat chance results. In fact, the results not only validate the objective and subjective procedures that were used to determine the seedings, but also provide a quantitative estimate of the success of those procedures. Moreover, our results (Table 1) show that the success ratio is correlated with the rank. With the exception of men's tennis, the won–lost percentage was best for the highest-ranked team (individual) and declined (but not monotonically) with lower rankings.

For the tennis rankings there is another measure of their predictive value. If the seeds were perfect predictors, all 16 of the ranked players would beat their unseeded opponents in each of the first three rounds of competition, and the competitors in the (fourth) round of 16 would all be seeded. This, however, is not what happened in this period; 47% of the male and 37% of the female players in the round of 16 were unseeded. Still, the rankings, while not perfect predictors, were informative. The four top-ranked male players were eliminated less frequently in the early rounds than were the other 12 seeded individuals[3] (Table 2). For the women tennis players the rankings were even more informative. Because of the dominance of Evert, Graf, Navratilova, and Seles, the two top-ranked individuals survived almost all the time and players ranked three and four made the round of 16 more than 80% of the time, the lower seeded women were less successful.

These data indicate that the rankings contain predictive value and that a forecast, 'the higher-ranked team will win', would have been correct a significant percentage of the time. By examining the relationship between rankings and the probability of winning a game or match, it might be possible to improve on this base rate prediction. Although the data are only presented below in Table 4, we also found that the relative frequencies of wins were positively related to the differences in ranks. This

Table 1
Winning percentage by rank among NCAA basketball teams and Grand Slam event tennis players

Rank	Basketball		Tennis	
	Men's	Women's	Men	Women
1	0.836	0.836	0.756	0.891
2	0.768	0.727	0.810	0.810
3	0.636	0.590	0.705	0.615
4	0.623	0.581	0.674	0.565
5	0.527	0.434	0.614	0.632
6	0.573	0.397	0.600	0.532
7	0.488	0.371	0.500	0.544
8	0.368	0.403	0.487	0.491
9	0.389	0.286	0.577	0.459
10	0.312	0.344	0.452	0.411
11	0.368	0.222	0.541	0.333
12	0.312	0.125	0.594	0.353
13	0.185	0.200	0.258	0.276
14	0.200	0	0.375	0.231
15	0.044	0	0.421	0.217
16	0	0	0.150	0.273
Unranked			0.163	0.111

[3] In the fourth and subsequent rounds, unseeded players defeat seeded individuals 11% of the times in men's tennis and 16% of the times in women's tennis. If this were representative of the relative performance in the early rounds, a female seeded player would have a 70% chance of reaching the fourth round; for males the comparable figure would be 59%. These percentages are consistent with the actual proportions of seeded players in the round of 16.

Table 2
Number of appearances (out of a possible 44) in the round of 16, men and women players by ranks 1–16

Rank	Men	Women
1	38	43
2	36	42
3	33	36
4	34	38
5	19	29
6	22	29
7	23	32
8	20	29
9	24	20
10	19	22
11	18	26
12	16	22
13	24	21
14	19	20
15	12	18
16	17	16

finding led naturally to the hypothesis that the probability of a higher ranked team beating a lower ranked one was a function of the difference in ranks. We, therefore, test the hypothesis that the probability of a higher-ranked competitor defeating an opponent is a function of the difference in ranks. Probit functions are used to test this hypothesis and to generate predictions. The statistical methodology is presented in the next section.

4. Statistical methodology

Although we analyze the four sports separately, the model is identical for each case. For every competition we derive a probit function that relates the outcome of a discrete event, winning or losing, to the difference of rankings. A separate probit model is fitted over the entire period for each of the competitions. We then ask the question: How well would these probits have predicted had they been used in real time to predict the outcomes of these sporting events? To analyze this question we use the technique of recursive regressions. Probits are fitted to a sample of past data, and the fitted function is used to forecast the next observation; then that next observation is included in the sample, the probit is refitted, and another prediction is generated, etc. The predictive accuracy of these probability forecasts is then assessed using techniques previously developed for evaluating the probability-of-precipitation weather forecasts.

4.1. Probit functions

The probit model is a statistical model relating the probability of the occurrence of discrete random events that take 0,1 values, such as winning or losing, to some set of explanatory variables. It yields probability estimates that the event will occur if the explanatory variables have specified values.

Specifically, let Y_i represent the outcome of an event, with $Y_i = 1$ indicating a win by a higher-ranked seed and $Y_i = 0$ otherwise. Then the probit can be specified as:

$$\text{Prob}[Y_i = 1] = \int_{-\infty}^{\beta' x_i} \phi(t)\mathrm{d}t,$$

where $\phi(t)$ is the standard normal distribution, x_i is a set of explanatory variables (such as the difference in ranks) for the ith observation, and β is a set of parameters to be estimated. (An alternative to the probit model is a logit model in which a logistic distribution replaces the standard normal distribution in the above equation. Results based upon the logit model are inconsequentially different from those based on the probit.)

4.2. Evaluation measures

The forecasts generated from the recursive probit regressions are probabilities. Measures specifically designed to evaluate probability forecasts are then used to determine how well these probit functions predicted the outcomes. For binary (the occurrence or non-occurrence of) events, the overall accuracy measure is the Brier score (mean square error):

$$B = \frac{\sum_{n=1}^{N} (r_n - d_n)^2}{N},$$

where r_n is the predicted probability that the event will occur on the nth occasion, and $d_n = 1$ if the event occurs on the nth occasion and zero otherwise. Smaller values of B indicate more accurate forecasts; a value of zero would indicate a perfect prediction. If rank had no explanatory power (i.e., if higher-ranked teams or players won only 50% of their games or matches), then the Brier scores would equal 0.25.

We also investigate the calibration or reliability of the forecasts (Murphy and Winkler, 1992, p. 437). The 0,1 range of values that the probability predictions might assume is first subdivided into t fractiles. Then the predicted probabilities of wins within a fractile are compared with the actual relative frequency of wins conditional on the predictions. Predictions are highly calibrated if the frequency of wins, conditional on the forecasts, corresponds to the predicted values.

5. Results

5.1. Probits

Table 3 presents the estimated coefficients of the probit equations for each sport and gender. All of the equations include as an independent variable the difference in seedings, $s_1 - s_2$, where s_1 is the seed of the higher-ranked team (player) and s_2 is the seed of the lower-ranked team (player). Alternative specifications of the difference in ranks variable (such as the square of the difference in ranks) do not improve either the fit of the equations or the quality of forecasts.

Since the tennis matches include unseeded players, we exclude all observations in which both players were unseeded (13.6% of men's matches and 5.9% of women's matches). When an unranked individual plays a ranked one, we arbitrarily assigned a value $s_2 = 17$ for the unranked player. We also include in the probit function a dummy variable (D) that equals one if the lower-ranked player is unseeded and zero otherwise. Thus, for an observation including an unseeded player, the linear function to be estimated is:

$$\alpha + \beta(s_1 - 17) + \gamma D,$$

where α is a constant term and $D = 1$. The advantage of this specification is that it is insensitive to the choice of the rank that is assigned to the unseeded player. For example, if 18 had been the assigned rank, then we would estimate

$$\alpha + \beta(s_1 - 18) + \delta D.$$

The estimated values of the constant term and the coefficient of the term measuring the difference in ranks would be unaffected, but the coefficient of the dummy variable would become $\delta = \gamma + \beta$ and the value of whole term would remain the same.

The *signs* of the coefficients provide information about the likelihood of a higher seed winning, while their magnitudes can be used to calculate the probabilities of winning. A significant negative coefficient on the independent variable would show that the probability of winning depends on the magnitude of the difference in ranks. The slope is negative because the better rank is the lower number.

In all of the estimated probits, the coefficient of the difference of ranks is negative and significant, thus indicating that the probability of winning an event does depend on the difference in the rankings of the competitors. The coefficients of the dummy variables for unseeded players are positive for both men and women when all tournaments are combined. This indicates that ranked players have a greater

Table 3
Coefficients and standard errors of probit regressions

Variable	Basketball		Tennis			
	Men	Women	Men			Women
			All tournaments	Wimbledon	Australia, France, United States	
Constant	−0.083 (0.098)	0.289 (0.110)	0.057 (0.117)	−0.449 (0.235)	0.218 (0.137)	0.239 (0.116)
Rank difference	−0.106 (0.015)	−0.125 (0.025)	−0.066 (0.015)	−0.146 (0.035)	−0.047 (0.017)	−0.090 (0.017)
Unseeded			0.230 (0.141)	−0.205 (0.360)	0.275 (0.155)	0.026 (0.155)
Brier score	0.180	0.164	0.170	0.160	0.170	0.140
n	660	476	570	141	429	621

probability of winning against an unseeded player for a given measured difference in ranks, although neither coefficient is statistically significant at conventional levels.

The relationship between the size of the difference in ranks and the probabilities of winning as calculated from the probit functions are presented in Table 4. The actual winning percentages associated with the differences in the rankings are in the same table. While the actual winning percentages are generally positively related to the magnitude of the difference in the rankings, they do not increase monotonically. In most cases there is a close correspondence between the predicted and actual frequencies.

We next tested whether the results for the tennis rankings would differ if the data for the Grand Slam tournaments were analyzed separately rather than pooled. We included interaction terms between dummy variables for the Grand Slam sites and the constant terms and independent variables in the probits and found that the coefficients for the women's tennis matches did not significantly differ among the four sites. For the men, however, the results for Wimbledon did differ significantly from the findings for the three other sites. Table 3 presents the probit equations separately for (1) Wimbledon and (2) the three other sites together. While the coefficients of the differences-in-ranks variables are significant in both equations, the coefficient of this variable in the Wimbledon equation is three times that of the equation for the other sites. The coefficient for the dummy variable for a match with an unseeded player is not statistically significant in the Wimbledon equation, but is significant at the 0.10 level in the equation for other events.

The differences in the relations between seedings and performance in Wimbledon compared to the other tournaments are shown in Table 5, which presents the predicted and actual frequencies by rank. The results are particularly striking if the difference in ranks is less than six. Weighted by the number of matches in which the difference in ranks is less than six, the predicted probability that a higher seeded player will win is only 0.48 and the actual relative frequency is 0.49. At the other three sites, the corresponding weighted averages of the predicted probabilities and actual frequencies of a higher-seeded player winning are 0.66 and 0.68, respectively. Differences are also prominent in the fraction of seeded players reaching the round of 16. While similar percentages of seeds 1–8 reach the round of 16 in Wimbledon and the other sites (67

Table 4
Predicted probabilities of winning and actual frequencies of the higher-ranked competitor winning by magnitude of difference in ranks

Difference in ranks	Basketball						Tennis					
	Men			Women			Men			Women		
	Pred.	Act.	# of events	Pred.	Act.	# of events	Pred.	Act.	# of events	Pred.	Act.	# of events
1	0.539	0.528	106	0.661	0.691	136	0.552	0.500	34	0.630	0.636	44
2	0.581	0.692	13	0.705	0.562	16	0.583	0.569	51	0.663	0.684	38
3	0.622	0.659	85	0.747	0.725	109	0.617	0.722	36	0.695	0.710	38
4	0.661	0.714	35	0.785	0.889	18	0.660	0.500	12	0.726	0.755	53
5	0.699	0.684	79	0.820	0.807	57	0.680	0.781	32	0.756	0.800	40
6	0.735	0.833	6	0.851	0.500	2	0.702	0.742	31	0.783	0.795	44
7	0.769	0.746	67	0.878	0.847	59	0.716	0.821	28	0.809	0.737	38
8	0.800	0.770	74	0.902	0.933	45	0.755	0.689	45	0.833	0.920	25
9	0.828	0.88	52	0.922	0.889	9	0.765	0.625	32	0.855	0.745	47
10	0.854	0.500	2	0.938		0	0.796	0.810	42	0.875	0.839	31
11	0.877	0.808	52	0.952	1.000	8	0.819	0.826	23	0.893	0.820	39
12	0.897	1.000	1	0.963	1.000	1	0.847	0.759	29	0.908	0.973	37
13	0.915	0.954	44	0.972	1.000	8	0.864	0.833	42	0.923	0.850	40
14	0.930		0	0.979		0	0.872	0.886	44	0.935	0.967	30
15	0.943	1.000	44	0.985	1.000	8	0.890	0.926	54	0.946	1.000	34
16							0.909	0.943	35	0.956	1.000	43

Table 5
Men's tennis by tour: predicted probability of the higher-ranked competitor winning by difference in ranks and actual winning proportion

Difference in ranks	Wimbledon			Australia, France, United States		
	Pred.	Act.	# of events	Pred.	Act.	# of events
1	0.372	0.333	9	0.604	0.560	25
2	0.437	0.500	14	0.636	0.595	37
3	0.495	0.667	9	0.665	0.741	27
4	0.553	0.000	3	0.710	0.667	9
5	0.586	0.600	10	0.712	0.864	22
6	0.642	0.714	7	0.724	0.750	24
7	0.701	0.600	5	0.730	0.870	23
8	0.722	0.400	5	0.763	0.725	40
9	0.795	0.667	6	0.769	0.615	26
10	0.812	0.929	14	0.792	0.750	28
11	0.838	1.000	5	0.809	0.778	18
12	0.863	0.625	8	0.833	0.810	21
13	0.892	1.000	8	0.850	0.794	34
14	0.921	1.000	12	0.853	0.844	32
15	0.939	0.846	13	0.870	0.951	41
16	0.953	1.000	13	0.893	0.909	22

versus 63%, respectively), only one-third of seeds 9–16 reach this round in Wimbledon compared to 45% in the other tournaments. Thus, the close positive relationship between rank and performance exhibited in women's tennis and the other men's tours does not apply to Wimbledon men's tennis, although it is not clear whether the Wimbledon results are attributable to the difficulty of predicting quality of play on grass surfaces or the substitution of judgment for computer rankings in determining seedings.

5.2. Predictions: recursive probit regressions

The probits presented above were estimated from data for the entire period. Because this information would not have been available in real time, we use the technique of recursive regressions to determine whether these probits would have provided useful real time forecasts. The probits are first estimated using only the 1985 data. Predictions are then generated for 1986, and the Brier score is calculated for these forecasts. Then the 1986 outcomes are added to the database from which the probits are estimated and the 1987 predictions are obtained. This procedure is then repeated sequentially in order to generate predictions for each year 1988–1995. The forecasts obtained from the probits were then compared with the predictions obtained from a base case. The base-rate probability forecast for each year was the historical winning percentage of higher-ranked teams (individuals) over lower-seeded competitors regardless of the difference in ranks. For example, in 1985 the higher placed men's basketball teams defeated their opponents 73.3% of the time. This number, 0.733, was used as the base rate for the 1986 forecasts. When the two years 1985 and 1986 were combined, this ratio was 0.725 and that figure was used as the 1987 base-rate probability, etc.

5.2.1. Results: basketball

For each year and for all forecasts taken together, the Brier scores for both the probit regressions and base-rate forecasts are reported in Table 6. The scores for both methods are quite low, indicating that the rankings contain information that would have been valuable in making ex ante forecasts about the outcomes of these basketball games. Merely predicting that a better-ranked team would defeat a lower-seeded competitor would have been accurate more than 70% of the time. For both basketball tourneys the Brier scores for all forecasts taken together are lower for the forecasts generated by the probits than those obtained from the base-rate predictions, as they are in 8 of 10 years for men's basketball and 7 of 10 years for women's basketball. (The probability of

Table 6
BRIER scores, basketball and tennis, recursive probits and base-rate forecasts

Year	Basketball				Tennis			
	Men		Women		Men		Women	
	Base rate	Probit	Base rate	Probit	Base rate	Probit	Base rate	Probit
1986	0.203	0.207	0.157	0.187	0.216	0.213	0.132	0.137
1987	0.210	0.190	0.124	0.112	0.164	0.160	0.154	0.138
1988	0.189	0.174	0.121	0.106	0.211	0.196	0.137	0.142
1989	0.211	0.175	0.164	0.160	0.204	0.177	0.137	0.139
1990	0.225	0.203	0.193	0.169	0.202	0.178	0.173	0.148
1991	0.182	0.185	0.205	0.227	0.170	0.134	0.173	0.147
1992	0.181	0.177	0.243	0.244	0.190	0.161	0.134	0.113
1993	0.166	0.160	0.188	0.182	0.154	0.175	0.154	0.157
1994	0.211	0.172	0.188	0.171	0.180	0.188	0.139	0.143
1995	0.180	0.165	0.143	0.126	0.158	0.151	0.147	0.149
All	0.196	0.180	0.174	0.169	0.185	0.173	0.148	0.141

eight successes in 10 trials if the underlying probability were 0.5 is statistically significant at the 0.054 level; the probability of seven successes in 10 trials is only significant at the 0.17 level.)

In order to measure calibration, we have grouped the entire set of predicted probabilities for the entire period, 1986–1995, into deciles (see Table 7). Within each decile the mean predicted probabilities that the higher-seeded competitor will win can be compared with the actual ratios of such outcomes. The men's basketball forecasts are well calibrated. While the women's basketball forecasts underpredict the outcomes in the 0.6–0.7 decile and overpredict in the next decile, in the other two deciles the predicted and actual outcomes are not too different.

5.2.2. Results: tennis

The Brier scores for the tennis predictions are also presented in Table 6. These scores are based upon recursive probits in which all tournaments are pooled. We also prepared separate forecasts for each tournament treated separately, and three alternatives in which France, Wimbledon, or both were treated separately. The Brier scores did not generally differ very much by choice of pooling. For both men's and women's tennis, the Brier scores for all years combined were lowest with all tours pooled and highest with each tour treated separately.

Our findings about the forecasts of the outcomes of tennis matches are similar to those obtained about the predictions of basketball games. Again, both the

Table 7
Calibration of recursive probit forecasts, basketball and tennis number of events and proportion of time better-ranked won

Predicted probabilities of better-ranked winning	Basketball				Tennis			
	Men		Women		Men		Women	
	# of events	Proportion	# of events	Proportion	# of events	Proportion	# of events	Proportion
0.40–0.49	10	0.70	0	[a]	9	0.54	0	[a]
0.50–0.59	116	0.55	0	[a]	74	0.55	12	0.59
0.60–0.69	159	0.67	59	0.76	94	0.67	102	0.66
0.70–0.79	149	0.78	127	0.61	124	0.77	140	0.75
0.80–0.89	102	0.83	132	0.83	183	0.85	143	0.86
0.90–0.99	64	0.97	130	0.89	33	0.91	169	0.93

[a]No prediction.

base-rate predictions and the probits have low Brier scores. While the Brier scores for the probit forecasts are smaller than the base-rate forecasts in 8 of 10 years for men's tennis, the base-rate forecasts for women's tennis are superior in only six of those years. The Brier scores for the probit forecasts for both the men's and women's competitions for all forecasts taken together are, however, lower than those for the base-rate predictions. Both men's and women's tennis forecasts are well calibrated (Table 7).

6. Conclusions

This paper has found that for both basketball tourneys and tennis matches the rankings provide forecasting information. This finding validates the procedures, based upon past performances and informed judgment, that have been used to generate those rankings. We then developed a model that used these rankings to predict the probability that one competitor would defeat another. We found that there was a significant relationship between the probability of winning and the difference of ranks, i.e. the probit functions have the appropriate and significant signs attached to the difference-in-ranks variable.

Our evaluation also demonstrates that, for each sport, when the yearly forecasts are combined, those obtained from recursively estimated probit functions are more accurate than the predictions generated by the standard of comparison. It is possible to test whether these results are statistically significant.

For each of the four tourneys both the recursive probits and the base-rate forecasts yielded 10 sets of predictions. The Brier scores for the probit predictions are lower in 27 of the 40 cases. The probability of observing this number of trials favoring the probits over the base-rate standard if the true probability were 0.5 (i.e. the forecasting methods were not significantly different) is only 0.02. On the basis of these results we can conclude that while knowledge of the ranks by itself is useful for forecasting, using the difference in ranks in a forecasting model improves the accuracy of the predictions.

Acknowledgements

We appreciate helpful comments from Robert Goldfarb, Ira Horowitz, Keith Ord, David Ribar, Christopher Snyder, two referees, and participants in ISF 1997 in Barbados.

References

Brown, W. O., & Sauer, R. D. (1993). Does the basketball market believe in the hot hand? Comment. *The American Economic Review*, *83*, 1377–1386.

Camerer, C. (1989). Does the basketball market believe in the hot hand?. *The American Economic Review*, *79*, 1257–1261.

Murphy, A. H., & Winkler, R. L. (1992). Diagnostic verification of weather forecasts. *International Journal of Forecasting*, *7*, 435–454.

Roberts, S. (1996). Agassi eludes the adult questions. *The New York Times*, August 6, C-9.

Sauer, R. D., Brajer, V., Ferris, S. P., & Marr, M. W. (1988). Hold your bets: another look at the efficiency of the gambling market for National Football League games. *Journal of Political Economy*, *96*, 206–213.

Schwertman, N. C., McCready, T. A., & Howard, L. (1991). Probability models for the NCAA regional basketball tournaments. *The American Statistician*, *45*, 35–38.

White, E. M., Dattero, R., & Flores, B. (1992). Combining vector forecasts to project thoroughbred horse race outcomes. *International Journal of Forecasting*, *8*, 595–611.

Biographies: Bryan L. BOULIER is Professor of Economics at The George Washington University. His research specialties are demography and applied microeconomics.

H.O. STEKLER is currently a Visiting Scholar in the Economics Department of George Washington University. He has written extensively in the area of forecasting, with particular emphasis on forecast evaluation.