International Journal of Forecasting 19 (2003) 257–270

www.elsevier.com/locate/ijforecast

Predicting the outcomes of National Football League games

Bryan L. Boulier*, H.O. Stekler

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

Abstract

Rankings have predictive value for determining the outcomes of basketball games and tennis matches. Rankings, based on power scores, are also available for NFL teams. This paper evaluates power scores as predictors of the outcomes of NFL games for the 1994–2000 seasons. The evaluation involves a comparison of forecasts generated from probit regressions based on power scores published in *The New York Times* with those of a naive model, the betting market, and the opinions of the sports editor of *The New York Times*. We conclude that the betting market is the best predictor followed by the probit predictions based on power scores. We analyze the editor's predictions and find that his predictions were comparable to a bootstrapping model of his forecasts but were inferior to those based on power scores and even worse than naive forecasts.

Keywords: Sports forecasting; Expert predictions; Bootstrapping

1. Introduction

In an earlier paper (Boulier & Stekler, 1999) we found that the rankings of teams in basketball tournaments and of tennis players in the Grand Slam events were good predictors of the outcomes of these competitions. Using statistical probit models, we concluded that there was predictive value in both the information that one competitor was considered superior to another and in the quantitative difference in the rankings. In many sports there is additional information contained in so-called 'power scores' that measure the relative abilities of teams based on objective criteria such as the team's performance and the strength of its own and its opponents schedules.[1] In a sport such as football, other objective criteria, such as the number of yards a team gained offensively versus the number of yards that its opponents gained, may be used.

There are a number of individuals who have developed power scores for the NFL teams and who have used these measures to predict the outcome of these games. The variables that are used to generate these power scores differ among individuals.[2] Some of these individuals present data on the Internet showing the percentage of times that their power-score measure correctly predicted the winner. These tabulations presented on the Internet suggest that power-scores have predictive value, because a majority of the individuals had success ratios in excess of 60%. Since we did not have access to the underlying power scores and the predictions of these

*Corresponding author. Tel.: +1-202-994-8088; fax: +1-202-994-6147.

E-mail address: mortile@gwu.edu (B.L. Boulier).

[1] Bassett (1997) and Glickman and Stern (1998) provide references for alternative methods of statistical estimation of power scores.

[2] Although the variables used in generating these power scores are sometimes published, the precise formulas for calculating these numbers are typically not presented.

0169-2070/01/$ – see front matter © 2001 International Institute of Forecasters. Published by Elsevier Science B.V. All rights reserved.
doi:10.1016/S0169-2070(01)00144-3

individuals, we could not systematically evaluate their performance.

Here, using the power scores published weekly in *The New York Times*, we undertake our own analysis of the value of power scores in predicting the winners of NFL games. *The New York Times* generates the power scores using the won–loss record of the teams, whether games were played at home or away, point margins of victory or loss, and the quality of opponents. Performance in recent games is given greater weight. The validity of generating power scores from these data is tested by comparing this approach with other methods of ranking teams. The approach used in assessing the predictive value of these power scores is similar to the one used in evaluating the accuracy of rankings in predicting the outcome of tennis matches and basketball games (Boulier & Stekler, 1999).

The *primary* purpose of the present paper is to determine whether the power scores (and the resulting rankings of professional football teams based on these power scores) published in *The New York Times* can forecast the winners of games in the National Football League (NFL). In evaluating the predictive value of these power scores, we compare the forecasts obtained from this measure with alternative methods for forecasting the outcomes of the NFL games. These alternatives include a naive model, the betting market, and the opinions of an expert sports editor. The betting market provides information both about the team that is expected to win and the number of points by which that team is favored. In this paper we will not examine the point spread, but will only ask whether the betting market can accurately predict which team will win an NFL game. No study has as yet addressed this question.[3]

Along with the power scores, *The New York Times* also publishes a sports editor's forecasts of the upcoming week's games. We can, thus, evaluate the forecasts of this expert, who not only has access to the power scores but also to outside information, to determine whether the power score predictions or the expert's forecasts are more accurate, and whether his predictions contain information that is not embodied in the power scores. We also use bootstrapping techniques to compare the sports editor's actual forecasts with a statistical model of his forecasts.

Section 2 describes our data. Section 3 compares forecasts of a probit model that estimates the probability that a higher-ranked team will win based on power scores and home-field advantage with forecasts made by *The New York Times* sports editor and the betting market. Section 4 examines in detail the forecasts of the sports editor and presents information on forecasts by other experts. Section 5 is a summary and presents our conclusions.

2. Data

During the sample period, the NFL had between 28 and 31 teams,[4] each of which played 16 games in a 17-week period from September to December. Our analysis is based on games played during weeks 6–17 of the 1994–1996 seasons, weeks 5–17 of the 1997–1999 seasons, and weeks 7–17 of the 2000 season, yielding 1212 observations. *The New York Times* publishes a power score that measures the relative abilities of each team. This measure was first published after the first 5 weeks of the 1994–1996 seasons, the first four games of the 1997–1999 seasons, and the first six games of the 2000 season. The power scores summarizes a team's relative performance in past games. It is based on the won–loss record of the teams, point margins of victory or loss, whether games were played at home or away, and the quality of opponents. Runaway scores are downweighted and performance in recent games is given greater weight. The top team is assigned a rating of 1.000 and relative power scores are used to

[3] Most studies that have examined the betting market sought to determine whether it was efficient. In this context, efficiency means that there are no systematic profit opportunities for beating the betting spread. Sauer (1998) surveys the betting market literature. In addition to the articles referenced by Sauer, Glickman and Stern (1998) and Avery and Chevalier (1999) also examine the betting spread.

[4] In 1994, which is the first year in our sample, there were only 28 teams. The number of teams increased to 30 in 1995 and to 31 in 2000. The teams are divided into two Conferences and within each Conference there are three Divisions. While each team plays 16 games, their opponents are selected on the basis of each team's performance in the previous year.

Table 1
Descriptive statistics[a]

Variable	Mean	Standard deviation	Minimum	Maximum
Difference in power scores	0.245	0.183	0.001	0.965
Difference in ranks	−10.528	7.086	−30	−1
Higher ranked team wins	0.608	0.488	0	1
Home team is higher ranked	0.483	0.500	0	1
Home team wins	0.611	0.488	0	1
NY Times editor picks higher ranked team to win	0.754	0.431	0	1
NY Times editor picks home team to win	0.549	0.498	0	1

[a] There are 1212 observations.

rank teams. Every week for the remainder of the season this measure is updated by taking into account the results of the previous week.

The power scores and resulting rankings are published several days prior to the next set of games. At virtually the same time, a sports editor publishes his predictions in *The New York Times*. The sports editor has the opportunity to use the quantitative indicator of the teams' relative abilities as well as other information that has not been incorporated into this measure. Such information might include data on injuries to key players, possible weather conditions, the home-field advantage, etc., as well as *subjective* views about the relative strength of teams. It is of interest to determine whether the expert's predictions are more accurate than those that would have been based solely on the rankings. Moreover, the forecasts based on the quantitative rankings and those obtained from the sports editor can both be compared with the predictions that come from the betting market. In this analysis we used the betting market predictions that were available either on the day before the game or on the day of the game itself.

Table 1 presents summary statistics for our data.

3. Results

3.1. Validity of The New York Times' power scores

Before we evaluate the forecasting accuracy of power scores, we present evidence verifying that the rankings derived from *The New York Times*' method were similar to those of other systems of estimating power scores. While we did not have rankings of other individuals for the time period covered by our primary data, we were able to gather such information for eight ranking systems for week 17 of the 1999 NFL season. Table 2 presents the Spearman rank correlation coefficients of the rankings of *The New York Times* with those obtained from the other sources. The results show that the rank correlation coefficients among the various ranking systems were indeed relatively high. The rank correlation coefficients of *The New York Times*' rankings with those published by others range from 0.83 to 0.97, thus establishing validity for the *Times*' method of generating power scores. The systems for ranking NFL teams employed by *The New York Times* and by Jeff Sagarin for *USA Today* are identical to those they use for ranking college teams. These systems are sufficiently respected that they are two of the computer ranking systems selected to determine the relative position of college football teams for deciding selections for end-of-season Bowl games.[5]

3.2. Forecasts: power scores and other methods

The next step is to actually evaluate the predictive accuracy of the power score forecasts for NFL

[5] *The New York Times*, Sagarin (*USA Today*), and Anderson-Hester (*Seattle Times*) computer ranking (i.e. power score) systems have been combined with rankings of teams derived from the Associated Press media poll and the *USA Today*/ESPN coach's poll to select places in the Bowl Championship Series since the initiation of this process of allocating college football bowl bids. For 1999 and 2000, five additional computer ranking schemes were added to the first three.

Table 2
Spearman rank correlations of power score ratings of 31 NFL teams by eight systems for week 17 of the 1999 NFL season[a]

	NY Times	Sagarin *USA Today*	*LA Times*	Huckaby	Kambour	Massey	Moore
Sagarin	0.837						
LA Times	0.885	0.850					
Huckaby	0.934	0.893	0.906				
Kambour	0.832	0.804	0.829	0.896			
Massey	0.957	0.873	0.928	0.976	0.884		
Moore	0.960	0.814	0.891	0.936	0.854	0.975	
Packard	0.970	0.821	0.875	0.913	0.849	0.937	0.944

[a] Websites: *The New York Times* (www.nytimes.com), Jeff Sagarin (www.usatoday.com/sports/sagarin/nfl00.htm), *The Los Angeles Times* (www.latimes.com), Stewart Huckaby (www.hometown.aol.com/swhuck/nfl.html), Edward Kambour (www.stat.tamu.edu/~kambour/NFL.html), Sonny Moore (www.members.aol.com/powerrater/nfl-foot.htm), Erik Packard (www.mesastate.edu/~epackard/nfl.html).

games. Since the rankings of teams correspond monotonically with the power scores, we do this by determining the 'within season' accuracy of forecasts obtained by predicting: 'the higher-ranked team will win'. Then the accuracy of these forecasts is compared with the accuracy of predictions obtained from: (1) the sports editor, (2) the betting market, and (3) a naive method of forecasting the outcome of the NFL games, i.e. the home team will win.

Table 3 indicates that over the entire period of our sample, forecasts based on power scores correctly predicted the outcomes about 60% of the time, indicating that models using power scores have predictive value. This result is consistent with the data that were obtained from the Internet. Moreover, predicting that the higher ranked team will win has a success rate slightly above that of the sports editor. However, naively predicting that the home team will win would have yielded a success rate nearly identical to the one obtained from the power score forecasts. Finally, the betting market made a higher percentage of correct predictions (of which team would win) over the entire period and in every year. In fact, the betting market was the most accurate forecasting method in every year of the sample.[6]

Using the data in Table 3, we tested the null hypothesis that these results could have been obtained by chance. The binomial distribution with

[6] In order to determine whether the proportions of games predicted correctly for each method of prediction varied across years, we estimated Poisson regressions for each method and tested whether the proportions were homogenous. For each method, we failed to reject the hypotheses that the proportions were homogenous over time at better than the $P = 0.60$ level for each case.

Table 3
Proportion of times that the team that is predicted to win actually wins by forecasting method, 1994–2000

Year	*N*	Power score: higher rank wins	*New York Times* sports editor	Naive: home team predicted to win	Betting market[a]	
					N	
1994	158	0.639	0.627	0.589	153	0.647
1995	170	0.535	0.571	0.612	167	0.629
1996	167	0.587	0.635	0.593	165	0.636
1997	183	0.623	0.557	0.607	178	0.674
1998	183	0.634	0.596	0.645	183	0.683
1999	189	0.624	0.577	0.614	189	0.667
2000	162	0.611	0.630	0.611	160	0.663
All years	1212	0.608	0.597	0.611	1195	0.658
Brier score—all years:		0.392	0.403	0.389		0.342

[a] If the spread was zero in the betting market, that observation was excluded. There were 17 such cases.

$P = 0.50$ was used to test the hypothesis. In all four cases (power scores, sports editor, betting market, and home team wins), this hypothesis was rejected at significance levels that were less than 0.01.

We next examined whether there was a significant difference between the won–lost records of the betting market and the naive ('home team will win') models.[7] We rejected the hypothesis that the two records were equal at the 1% level of significance. That is, there is less than a 1% chance that the betting market would have picked the winner as frequently as it did if the true probability of making a correct prediction were identical to the naive model (i.e. $P = 0.611$).[8]

Although the percentage of accurate forecasts is adequate for comparing the alternative forecasting methods, in Table 3 we present another statistic, the Brier Score (Brier, 1950; Schmid & Griffith, 1998), which will also be used in our subsequent analyses. The Brier Score, which is specifically designed to evaluate probability forecasts, is defined as:

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

where $d_n = 1$ if the event occurs on the nth occasion and equals 0 otherwise, and r_n is the predicted probability that the event will occur on the nth occasion. Predicting that the higher ranked team will win can be interpreted as a forecast that the event will occur with probability 1. If the forecasts are always accurate, QR is 0; if the predictions are correct 50% of the time, the value of the statistic equals 0.5. Predictions that are always wrong produce $QR = 1$. Whenever the values of r_n are constrained to equal 1 or 0, the Brier Score equals the proportion of wrong predictions. The Brier Score is equivalent to the mean square error of the probability forecasts. The Brier Scores range from 0.342 for the predictions of the betting market to 0.403 for the predictions of the sports editor.

3.3. Probits

Our analysis of the forecasts derived solely from the rankings based on power scores did not take into account either the quantitative difference in the ranks or the difference in the power scores. In this section, we estimate a probit model designed to predict the probability that the higher ranked team will win. The explanatory variables will either be the difference of ranks or the difference in power scores from which the rankings are derived. We can then determine whether it is possible to improve upon the performance of the forecast: 'the higher ranked team will win'. Thus it will be possible to determine whether the difference in ranks and/or power scores of two teams contains information beyond merely knowing that one team has a higher rank than the other.

A probit statistical model is a statistical model relating the probability of the occurrence of discrete random events (Y_i), which take 0, 1 values such as winning or losing, to some set of explanatory variables. It yields probability estimates of the event occurring if the explanatory variables have specified values. Specifically, if we let $Y_i = 1$ for a win and $Y_i = 0$ otherwise, then the probit can be specified as:

$$\text{Prob}[Y_i = 1] = \int_{-\infty}^{\beta' x_i} \phi(z)\,\mathrm{d}z$$

where $\phi(z)$ is the standard normal distribution, x_i is a set of explanatory variables (such as the difference in ranks) for game i, 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.[9])

[7] The model is not *completely* naive, because it would have required an individual to at least learn that there is a home-field advantage.

[8] On the other hand, there is only a 17% chance that the sports editor could have done as poorly as he did if the true probability of picking the winner were $P = 0.611$. In all cases, a one-tail test is used.

[9] Predicted probabilities of winning derived from logit regressions are nearly identical to those obtained from probit regressions. For example, using the difference in ranks (i.e. rank of higher ranked team minus rank of lower ranked team) as an independent variable in the probit regression, the predicted probability of the higher ranked team winning is 0.514 when the difference in ranks is -1. This probability is 0.782 when the difference in ranks is -30 (see Table 5). The corresponding probabilities derived from a logit regression are 0.514 and 0.778, respectively.

Two sets of probits were estimated using pooled data for the 5-year period. In one case, the independent variable was the difference in rankings between the two teams. In the other, the difference in power scores was used as the explanatory variable. To account for the possibility of a home-field advantage, a dummy variable was added as an additional explanatory variable in both sets of probits. The dummy variable has a value 1 if the higher ranked team is playing on its home field and is 0 otherwise. Since the better ranked team has a lower number, a negative coefficient of the difference in rankings variable would show that the probability of the higher ranked team winning is larger the greater is the difference between its rank and that of its opponent. The coefficient of the variable measuring the difference in power scores is expected to be positive. A positive coefficient on the dummy variable would indicate that the probability of the higher-ranked team winning increases if it is playing at home.

Table 4 presents the coefficients of the probits. For the entire sample period, the coefficients of the difference in ranks/power score variables have the expected signs and are statistically significant in all cases. Thus, the difference in ranks/power scores is a significant factor in determining the probability that the higher-ranked team will win. The home-field dummy also had the expected sign and was statistically significant in all cases, confirming the often stated views of football coaches that playing at home increases a team's chances of winning.[10] The pseudo-R^2s presented in Table 4 are, however, quite low. This finding indicates that, while difference in ranks/power scores and home-field advantage are systematically related to a team's probability of winning, many factors not captured in the regression (such as injuries to key players, whether a touchdown pass is on target or slightly too long or too short, whether infractions of the rules are detected or not, etc.) affect outcomes on 'any given Sunday'.[11]

We also tested whether the coefficients of the constant term, the difference in ranks or power scores, or the home-field dummy in the probit equation differed by year or by week of the season. In all cases, we rejected decisively the hypotheses that coefficients differed by year or week either when the coefficients were considered individually or as a

Table 4
Coefficients of probits based on differences of ranks and differences of power scores, without and with a dummy variable for home team advantage

Independent variable	Dependent variable: higher ranked team wins			
	(1)	(2)	(3)	(4)
Constant	0.0098 (0.0655)	−0.2949* (0.0759)	0.0289 (0.0614)	−0.2750* (0.0723)
Difference in ranks	−0.0256* (0.0053)	−0.0273* (0.0054)		
Difference in power scores			1.0246* (0.2087)	1.0905* (0.2130)
Home team is higher ranked dummy		0.6263* (0.0753)		0.6268* (0.0754)
Pseudo R^2	0.01	0.06	0.02	0.06

Notes: standard errors in parentheses. An asterisk indicates that the coefficient is statistically significant at the 0.01 level. There are 1212 observations.

[10] See Vergin and Sosik (1999, p. 23) for a brief review of the literature on the sources of the home-field advantage.

[11] A low pseudo-R^2 does not mean that rank is not a powerful predictor of outcomes of games on average. Suppose, for example, there are 200 games. Assume that in 100 of these games there is a big difference in ranks of opponents and in 100 there is a small difference in ranks. Suppose further that the higher ranked team wins 75% of games when there is a 'big difference' in ranks and 55% of games when there is a 'small difference' in ranks. For this sample of games, a probit regression estimates that the marginal effect of a dummy variable denoting a 'big difference' in ranks is 0.20 and is statistically significant at the 0.003 level. Yet, the pseudo-R^2 is only 0.034.

set. For example, a test of the hypothesis that the coefficients of the differences in ranks equations were identical across years yields a χ^2 of 9.30 with 12 degrees of freedom ($P = 0.68$).

Table 5 presents the relationships between the size of the difference in ranks and (1) the actual proportion of wins and (2) the predicted probability of winning derived from the probit regression using only the difference in ranks (column 1 in Table 5). The predicted probabilities imply that a team that is ranked one position higher than its opponent has only a slightly better than even chance of winning, while a team that is ranked 30 positions above its opponent should win more than three-quarters of the time. The actual proportion of games won by the higher ranked team does not increase monotonically with the difference in ranks. Some of this variation is the result of the small number of games corresponding to various rank differences. If games are combined into larger categories, the relationship between the differences in rank and the proportion of wins becomes clearer. For example, the difference in ranks is three or less in 13% of games included in the sample. For these games, the weighted average of the predicted percentage of games won by the higher ranked team is 52% and the weighted average of the actual percentage is 54%, where the weights are the number of games in each rank difference category.

Table 5
Predicted and actual probabilities of winning by difference in ranks, 1994–2000

Difference in ranks	Predicted probability of higher ranked team winning	Actual proportion of wins by higher ranked team	
		Number of games	Proportion of wins
30	0.782	2	1.000
29	0.774	6	1.000
28	0.767	9	0.778
27	0.759	9	0.667
26	0.751	11	0.636
25	0.742	16	0.813
24	0.734	16	0.750
23	0.725	17	0.824
22	0.717	25	0.760
21	0.708	22	0.909
20	0.699	25	0.720
19	0.690	25	0.720
18	0.681	35	0.714
17	0.672	42	0.548
16	0.663	40	0.625
15	0.653	47	0.596
14	0.644	43	0.605
13	0.634	43	0.605
12	0.625	53	0.509
11	0.615	60	0.617
10	0.605	63	0.667
9	0.595	56	0.518
8	0.585	46	0.522
7	0.575	68	0.603
6	0.565	69	0.507
5	0.555	65	0.662
4	0.545	64	0.594
3	0.535	80	0.575
2	0.524	69	0.565
1	0.514	86	0.477

At the other end of the spectrum, 13% of games involved differences in ranks of 20 or more. For these games, the weighted average of the predicted percentages of games won by the higher ranked team is 73% and the weighted percentage of actual games won is 78%. Since the rankings are based upon power scores, these results show that the scores have predictive value.

In addition, the coefficient of the dummy variable indicating whether the home team is higher ranked (column 2 in Table 4) reveals the powerful effect of home-field advantage. The probit equation including the home-field dummy along with the difference in ranks implies that, when the difference in ranks is only -1 (that is, the higher ranked team is ranked only one position higher than its opponent), the higher ranked team has a probability of winning of 0.64 if it is playing on its home field compared to only 0.39 if it is playing away. When the difference in ranks is -30, the higher ranked team has a predicted probability of winning of 0.87 on its home field compared to 0.70 if playing away.

The results described above indicate that rankings and power scores are strongly associated with the outcomes of games. It is, however, necessary to determine whether the rankings and power scores would also have predictive value if they were the basis of forecasts made in real time. We use recursive probit regressions to examine this issue.

3.4. Predictions: recursive probit regressions

The probits that were presented above were estimated from data for the entire period. This information would not have been available in real time. We, therefore, use the technique of recursive regressions to determine whether these probits would have provided useful forecasts in real time. As an example of this procedure, consider the year 1994. The probit equations shown in Table 4 are estimated first with the data available for weeks 1–6. These probits are then used to predict the outcomes of the week 7 games. After the week 7 games have been played, the probits are updated and predictions are generated for week 8, etc. This procedure is then repeated in order to generate these probability predictions for the remainder of each football season. A similar process is applied to the data for each year of the sample. Then, for each year, the Brier Scores for the probability forecasts obtained from the probits are calculated. These Brier Scores are presented in Table 6. The Brier Score obtained from aggregating the predictions generated separately for each year is also presented.

Table 6
Brier scores for recursive probit regressions and for two naive models, *The New York Times* sports editor, and the betting market, 1994–2000

Year	*N*	Recursive probit regressions					Naive models		*New York Times'* sports editor	*N*	Betting market
		Difference in ranks	Difference in ranks plus home team dummy	Difference in power scores	Difference in power scores plus home team dummy		Higher ranked team picked	Home team picked			
		Probability forecasts				Win/loss					
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1994	146	0.236	0.238	0.237	0.239	0.411	0.363	0.418	0.363	141	0.362
1995	157	0.257	0.259	0.255	0.257	0.408	0.471	0.408	0.420	155	0.381
1996	155	0.242	0.235	0.243	0.235	0.387	0.419	0.394	0.361	153	0.373
1997	170	0.238	0.229	0.237	0.231	0.359	0.382	0.406	0.459	165	0.339
1998	171	0.225	0.209	0.226	0.209	0.316	0.357	0.357	0.404	171	0.322
1999	175	0.242	0.233	0.243	0.234	0.389	0.365	0.377	0.400	175	0.331
2000	148	0.254	0.247	0.241	0.236	0.419	0.399	0.392	0.385	146	0.356
All years	1122	0.242	0.235	0.240	0.234	0.382	0.393	0.392	0.400	1106	0.351

In order to evaluate the predictive value of these probit forecasts, they are compared with the predictions obtained from (1) *The New York Times* sports editor, (2) the betting market, and (3) two naive base cases for the relevant years.[12] The first naive forecast was that the higher-ranked team would always defeat its opponent regardless of the difference in ranks. The second naive forecast is that the home team will win regardless of whether it is the higher-ranked team or not. The Brier Scores for these naive forecasts and for the predictions of the betting market and the sports editor are also presented in Table 6.

Consider first the recursive probit predictions of the *probability* that the higher ranked team would win (columns 3–6 of Table 6). For each year, there is very little difference in the Brier Scores of the recursive probits that use the difference in ranks vs. those that use the difference in power scores. The inclusion of the home-field dummy, however, somewhat reduces the Brier Scores, especially in 1998, thus indicating that it adds predictive value.

As judged by the Brier Scores, the probability predictions of the recursive probit regressions are superior to those obtained from the other approaches in every year. But, it must be remembered that the forecasts made by the naive models, the sports editor, and the betting market are unconditional predictions of wins and losses—not forecasts of the probability of winning. If the likelihood of a victory by a team in a given game exceeds the probability of a loss by only a small margin, a forecasting procedure that predicts the probability of a win or loss will have a lower Brier Score than one constrained to predict victory or loss.

Thus, in order to provide a more valid comparison of the rolling probit forecasts with those of other methods, we have used the forecasts of the model with power scores plus the home-team dummy to generate forecasts of wins or losses, predicting a win if the forecasted probability of a win exceeds 0.5 and a loss otherwise. These results are shown in column 7 of Table 6. If we compare the Brier Scores for wins and losses, the betting market (column 12) has the lowest Brier Score for all games combined and is the lowest in each year except 1996, when the sports editor outperforms the betting market. The forecasts of the recursive probit model have the second lowest Brier score for all games combined and this method outperforms the naive models in most years and the sports editor in 4 of the 7 years. The sports editor has the highest (worst) Brier score for all years combined. The naive approach of picking the home team to win is superior to the sports editor's predictions in 4 of the 7 years.

4. Statistical model versus expert opinion

4.1. Does the sports editor provide useful information?

There is a large literature analyzing whether the predictions obtained from informed judgment are more accurate than the forecasts derived from statistical models. A recent survey (Bunn & Wright, 1991) of previous studies that examined the role of judgment in forecasting concluded that the use of judgment in a real world context has value (p. 512). On the surface, our results (presented in Table 3) seem to suggest the opposite, because the ranking system based on power scores had a higher percentage of correct predictions about the outcome of the NFL games than did the sports editor. This result appears surprising, since the sports editor had not only knowledge of the power scores, but presumably also useful current information (such as injuries to quarterbacks or key players) relevant to forecasting game outcomes that would not be incorporated in the rankings or power scores. In this section, we investigate the sports editor's forecasts in more detail.

We first estimated two probit regressions to determine whether the sport editor's forecasts reflected publicly available information such as the power scores and home-field advantage. In the first probit regression (Table 7, row 1), the dependent variable is a variable that takes on the value one if the sports editor forecasts the *higher ranked team* to win and is zero otherwise. The independent variables in that regression are (1) the difference in power scores between the higher ranked and lower teams and (2)

[12] The results presented in Table 6 differ from those in Table 3. The Table 6 results omit the first week of each season covered by our data (e.g. week 6 in 1994) as the recursive probit procedure does not generate a forecast for the first week.

Table 7
Probit regressions examining *The New York Times* editor's picks, 1994–2000

Row	Dependent variable	Independent variables						
		Constant	Power score difference: higher ranked team minus lower ranked team	Home team is higher ranked	Editor picks higher ranked team to win	Power score difference: home team minus away team	Editor picks home team to win	Pseudo-R^2
(1)	Editor picks higher ranked team to win	−0.1525 (0.0806)	3.0646* (0.2900)	0.4337* (0.0846)				0.12
(2)	Higher ranked team wins	−0.3289* (0.0861)	1.0138* (0.2229)	0.6148* (0.0761)	0.1040 (0.0903)			0.06
(3)	Editor picks home team to win	0.2172* (0.0424)				3.3080* (0.1799)		0.29
(4)	Home team wins	0.2598* (0.0626)				1.1117* (0.1544)	0.0929 (0.0899)	0.06

Notes: standard errors are in parentheses. An asterisk indicates significance at the 0.01 level. There are 1212 observations.

the home-field dummy that takes on the value one if home team is higher ranked and zero otherwise. The dependent variable in the second regression (Table 7, row 3) takes on the value one if the sports editor picks the *home team* to win and is zero otherwise and the independent variable is the difference between the power scores of the home team and the away team. In both regressions, the coefficients of the independent variables have the expected signs and are statistically significant at the 0.01 level. This indicates that the editor uses this information in making his predictions.

We then tested whether the sports editor's predictions contain useful information not embodied in the power scores or home-field advantage. We followed the approach of Forrest and Simmons (2000) who analyzed outcomes of English soccer games and found that only one of three newspaper tipsters' predictions contained useful information not derivable from readily observable past performance (such as goals scored in the previous five home and away games). To measure the information content of the sports editor's forecasts, we also estimated two probit regressions. The first (Table 7, row 2) regresses whether or not the higher ranked team won on (1) the difference in power scores, (2) the home-field dummy, and (3) a dummy variable equaling one if the sports editor picked the higher ranked team to win and zero otherwise. The second (Table 7, row 4) regresses whether the home team won on (1) the difference in power scores between the home and away teams and (2) a dummy variable equaling one if the sports editor picked the home team to win and zero otherwise. In both cases the coefficients of the sports editor's picks are positive, but neither of the coefficients is statistically significantly different from zero at the 0.20 level. If the sports editor had provided useful information, the 'editor dummy' should have been positive and significant.

On the hypothesis that the *The New York Times* sports editor would have special knowledge about the two New York teams, we investigated whether his predictions of games involving the Giants and Jets provided information not generally available otherwise. We failed to find any evidence that his predictions of these games were more accurate than his forecasts of other games. For example, he successfully picked the winning team in 54.7% of the 159 games involving either the Giants or Jets vs. 60.5% of the 1053 games not involving these two teams. In probit regressions specified as in Table 7 (columns 2 and 4) but estimated only for the sample of games involving the Giants and Jets, the coefficients of the sports editor's choices of the higher ranked team to win or the home team to win were *negative* although not significantly different from zero at greater than the 0.50 level.

In short, we find no evidence that the sports editor adds any information that aids in prediction of football game outcomes beyond the information contained in the power scores and home-field advantage.

4.2. Bootstrapping the sports editor

Since the probit regression in Table 7 (column 1) indicated that power scores and home-field advantage were statistically significant predictors of the sports editor's choice of whether the higher ranked team would win, we also examined whether a statistical model of his choice would provide a superior forecast to that supplied by the sports editor. The 'bootstrapping' literature, nicely reviewed by Bunn and Wright (1991, pp. 505–556), has investigated the effectiveness with which statistical models of experts' judgments can represent the predictions of the experts. In this literature, experts' (or judges') decisions are statistically related to information used in an intuitive way by the experts. Then comparisons are made between predictions derived from the statistical models and those provided by experts. Whether statistical models can outperform experts is an open question. On the one hand, statistical models relating decisions by an expert to information used intuitively by the expert might provide superior predictions, because the statistical models are less subject to random error. On the other hand, simple (generally linear) models might not capture adequately the decision processes of the experts and, in real world situations, 'human judgment can theoretically improve on statistical modelling by recognising events that are not included in the model's formula and that countervail the actuarial conclusion' (Bunn and Wright, 1991, p. 506). In their review of past studies, Bunn and Wright (1991, p. 506) conclude:

> Although early studies of linear modelling in clinical settings showed evidence that the model of the judge outperforms the judge on which the model was based, the evidence of poor performance in experimental environments which are artificial in terms of the information available, in principle, to the forecaster, does provide a seriously restricted valuation of the quality of experienced judgment. In the real-world studies, where the forecasting process is less stable and less routine, with a need to incorporate special peripheral information, then the experiments have outperformed bootstrapping models.

Our application of the bootstrapping approach uses recursive probit regressions. We illustrate the procedure using data for 1994. We first estimated a probit equation in which the dependent variable was whether or not the sports editor forecasted the higher ranked team to win in week 6. The independent variables were the difference in power scores between the higher and lower ranked teams and the home-field dummy variable. This probit was then used to forecast the outcomes of games in week 7 of the season based on the power scores then available. As previously explained, the probit regressions are then updated for the rest of the season.

The results of this exercise are summarized in Table 8. The sports editor forecasts whether the higher ranked team wins or loses, but the bootstrap procedure provides estimates of the probability that the higher ranked team will win. Hence, we have converted the forecasted probabilities to 'win or lose' values by assigning a win if the forecasted probability exceeds 0.5 and a loss otherwise. The Brier Scores of the bootstrap forecasts of wins or losses (shown in column 2) are higher than the Brier Scores of the actual forecasts of the sports editor (column 1) for the years 1994–1996 and 2000 but much lower for 1997–1999. For example, in 1997 the bootstrap model of the sports editor correctly predicted the outcomes of 62% of games compared to his actual 54% success ratio. For all games combined, the Brier Score based on bootstrap estimate (0.388) is slightly lower than the Brier score for the actual forecast (0.404). We conclude that the statistical model of the sports editor's forecasting procedure yields slightly superior forecasts to the actual forecasts in a 'real world' situation for a large sample of predictions, but the difference is quite small.

4.3. Forecasts of other experts

Our comparison of predictions of outcomes of football games using statistical modeling with the actual forecasts of the sports editor indicates that the model-based forecasts are slightly superior to those of *The New York Times* sports editor, who has information not embodied in summary measures of

Table 8
Brier Scores of *The New York Times* editor's picks compared with bootstrap estimates

Year	*N*	(1) *New York Times* sports editor's picks	(2) Bootstrap forecasts of win or loss
1994	146	0.363	0.404
1995	157	0.420	0.446
1996	155	0.361	0.381
1997	170	0.459	0.382
1998	171	0.404	0.351
1999	175	0.400	0.366
2000	148	0.385	0.392
All years	1122	0.400	0.388

teams past performance and home-field advantage. However, this sports editor is merely a sample of one and it is possible that other experts might provide superior estimates to either the forecasts of the statistical models we have estimated or those of this sports editor. Indeed, it appears that the forecasts of *many* individuals who have either been players or coaches in the NFL are at least as accurate as the predictions derived from the ranks/power scores model. These results are presented in Table 9. (We cannot include these individuals in our analysis because the record of their predictions covers the entire NFL season while we only make comparisons for those weeks for which the power scores were available. The week by week record of predictions of these sports commentators was not available.) In fact, one may argue that the betting market, which represents the collective judgment of a large number of well-informed individuals, is the best measure of the judgmental forecast.[13] If the bettors judge that the point spread on the favored team is inappropriate,

Table 9
Accuracy of judgmental forecasts of outcomes of professional football games: *The New York Times* sports editor and four former NFL players/coaches

Year	*New York Times* sports editor[a]		Chris Collinsworth		Jerry Glanville		Nick Buoniconti		Len Dawson	
	No. of games	Percent correct	No. of games	Percent correct	No. of games	Percent correct	No. of games	Percent correct	No. of games	Percent correct
1994	158	62.7	–	–	–	–	–	–	–	–
1995	170	57.1	–	–	–	–	–	–	–	–
1996	167	63.5	251	67.3	251	57.0	251	65.3	251	63.8
1997	183	55.7	251	66.1	251	57.4	251	60.6	251	61.4
1998	183	59.6	251	68.9	251	61.4	251	69.7	251	67.7
1999	189	63.0	–	–	–	–	–	–	–	–
2000	162	59.7	–	–	–	–	–	–	–	–

Source: *The New York Times* and www.combellsvil.edu/nfl/nfl98.txt
[a] The number of games included weeks 6–17 of the regular season in 1994–1996 and weeks 5–17 in 1997–1998.

[13] The opening betting line or point spread is set by a number of informed football experts, who use judgment, working in conjunction with the Las Vegas odds-makers. This opening line is then adjusted to reflect the betting actions of the market participants. See Gandar, Dare, Brown and Zuber (1998) for an examination of betting-line adjustments in professional basketball. In an experiment conducted at the University of Indiana in which faculty and students predicted outcomes of professional and collegiate football games in 1966 and 1967, Winkler (1971, pp. 682–683) found that consensus forecasts outperformed the average of individual subjects' forecasts.

they will bet accordingly. Their actions will change the spread and may even change which team ends up being favored. If this view is correct, then the judgmental forecast, represented by the betting market, was more accurate than the statistical model based on power scores because it made the highest percentage of correct predictions about the outcome of the NFL games.

5. Summary and conclusions

We have questioned whether quantitative measures of relative team performance (power scores) can predict the outcomes of professional football games. *The New York Times*' power scores and the rankings that were based on those numbers were generally consistent with similar measures obtained from other sources.

The accuracy of forecasts made using these rankings were then compared with the records of (1) a sports editor who can use a variety of techniques, including the use of judgmental adjustments, (2) naive models; and (3) the betting market. Predictions based on power scores were slightly more accurate than those of the sports editor, but inferior to the betting market and even a naive model (i.e. the home team wins). We must emphasize that the naive model (i.e. the home team wins) requires a minimum of sports knowledge and forecasting expertise. Nevertheless, it out-performed the sports editor by a small margin for all years combined.

Probit regressions revealed that differences in ranks/power scores and home team advantage were statistically significant and quantitatively important determinants of whether the higher ranked team would win. Probability predictions were then made from recursive probit regressions, which is the way information on differences in team quality would be used in real time. The predictions from the recursive probit regressions based on rankings/power scores and including a dummy for the home team were superior to naive forecasts and the sports editor, suggesting that this method of predicting outcomes of NFL games has predictive value. However, the recursive probit predictions were inferior to forecasts based on the betting market.

We examined in some detail the predictions of *The New York Times* sports editor to determine whether this expert possessed useful forecasting information apart from that embodied in the power scores and home-field advantage. We found that his predictions of whether a higher ranked team would win or whether the home team would win did not improve upon forecasts based only on the power scores and home-field advantage. We also found that the statistical (bootstrapping) models of the sports editor's forecasts of whether the higher ranked team would win slightly outperformed his actual forecasts. Although the forecasts of statistical models using power score data were more accurate than the predictions obtained from a *New York Times* sports editor, there were other sports commentators whose forecasts were better than the model's predictions.

Overall, we conclude that the information contained in the betting market is the best predictor of the outcomes of NFL games. Nevertheless, an objective measure such as the power scores published in *The New York Times* is also informative. We have not tested whether other objective measures, such as the number of yards gained, the number of yards given up, the give-away take-away turnover ratio, etc., might be even more informative. (This is obviously an appropriate subject for further research.)

We also examined the subjective predictions of one sports editor and found that his predictions were comparable to a bootstrapping model of his forecasts but were inferior to those based on the objective method and even worse than naive forecasts. Nevertheless, the small sample size, conflicting results, and narrow differences among methods of prediction, did not permit us to reach a definitive conclusion about the relative merits of statistical model and expert approaches for forecasting the outcomes of professional football games. This question can only be answered with further research with a larger sample of experts and additional objective measures for evaluating the relative strength (or abilities) of NFL teams.

Acknowledgements

We appreciate the helpful comments of Keith Ord, an Associate Editor, and the two referees.

References

Avery, C., & Chevalier, J. (1999). Identifying investor sentiment from price paths: the case of football betting. *Journal of Business*, *72*, 493–521.

Bassett, Jr. G. W. (1997). Robust sports ratings based on least absolute values. *The American Statistician*, *51*(2), 99–105.

Bunn, D., & Wright, G. (1991). Interaction of judgmental and statistical forecasting methods: issues and analysis. *Management Science*, *37*, 501–518.

Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors? *International Journal of Forecasting*, *15*, 83–91.

Brier, G. W. (1950). Verification of weather forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Forrest, D., & Simmons, R. (2000). Forecasting sport: the behavior and performance of football tipsters. *International Journal of Forecasting*, *16*, 317–331.

Gandar, J. M., Dare, W. H., Brown, C. R., & Zuber, R. A. (1998). Informed traders and price variations in the betting market for professional basketball games. *The Journal of Finance*, *53*(1), 385–401.

Glickman, M. E., & Stern, H. S. (1998). A state-space model for national football league scores. *Journal of the American Statistical Association*, *93*, 25–35.

Sauer, R. (1998). The economics of wagering markets. *Journal of Economic Literature*, *36*(4), 2021–2064.

Schmid, C. H., & Griffith, J. L. (1998). Multivariate classification rules: calibration and discrimination. In Armitage, P., & Colton, T. (Eds.), *Encyclopedia of Biostatistics*, vol. 4. New York: John Wiley, pp. 2844–2850.

Vergin, R. C., & Sosik, J. J. (1999). No place like home: an examination of the home field advantage in gambling strategies in NFL football. *Journal of Economics and Business*, *51*, 21–31.

Winkler, R. L. (1971). Probabilistic prediction: some experimental results. *Journal of the American Statistical Association*, *66*, 675–685.

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

H.O. STEKLER is a Research Professor in the Economics Department of The George Washington University. He has written extensively on many aspects of forecasting with a particular emphasis on forecast evaluations.