Replication of:

Are difference in ranks good predictors for Grand Slam tennis matches?

(Julio del Corral & Juan Prieto-Rodríguez (2010))

By: Lexi Bender, Yanni Konstantinidis, James Ledoux, Carly Munnelly

Maxwell May 4, 2016

Introduction

Forecasting the outcome of competitive sporting events has long been an intriguing puzzle, due in part to the growing popularity of online sports betting (Leitner 2008). Sporting events are distinct from completely randomized events such as lotteries, lending themselves to modeling and prediction. The uses of rankings and ratings have improved prediction in competitive domains ranging from chess (Elo 2008) to international soccer matches (Dyte & Clark 2000). Due of the popularity of the topic, there is an extensive literature covering the available methods used to predict the result of a particular match, tournament, or league, with much debate surrounding the hierarchy of said techniques. The features examined in these models can essentially be split into three main categories: the past performance of a player, a player's physical characteristics, and unique match characteristics. These descriptive traits develop rankings which in turn have a relationship to the relative frequency of wins a competitor earns. Utilizing this insight, this paper looks to mimic the statistical methodology used by Corral and Prieto-Rodriguez (2010) in *Are differences in ranks good predictors for Grand Slam tennis matches?* And to expand upon this work by extending its time horizon and comparing its rankings to those from other potential models.

The men's and women's Grand Slams consist of four major tennis tournaments: the Australian Open, French Open, US Open, and Wimbledon. The contest kicks off with 128 entrants, of which only 16 are seeded. The format of the competition is such that the seeded participants only play unseeded contestants until the fourth round, creating an intriguing but small opportunity for an underdog run. While seedings may not be available for comparison in the early rounds, the ATP and WTA do release rankings which can serve as direct comparisons between perceived player abilities. The goal of this paper is to examine the extent to which these official rankings and others can be used to predict Grand Slam match outcomes.

Are differences in ranks good predictors for Grand Slam tennis matches?

Corral and Prieto-Rodriguez tested whether or not discrepancies in rankings between individual players were explanatory predictors for outcomes in Grand Slam tennis competition in their 2010 paper *Are differences in ranks good predictors for Grand Slam tennis matches?*, published in The International Journal of Forecasting. They utilize the WTA and ATP men's and women's Grand Slam tennis match data from 2005 to 2008 to estimate three different probit models. The paper tests the impacts of match and player characteristics on the binary dependent variable of whether the higher-ranked player wins. The match characteristics denote the tournament and round, and the player characteristics are split into two components capturing a player's past performance and physical characteristics.

In the first model, all three explanatory variables (past performance, physical characteristics, and match characteristics) are included. In the second and third models respectively, past performances and physical characteristics are omitted. The predictive accuracies of these models are then assessed by computing Brier scores and comparing the predicted outcomes of matches observed outcomes from 2005 to 2008. Finally, the predictive capability of the models is measured by running these same tests on out-of-sample data from the 2009 Australian Open.

The conclusion of Corral and Prieto-Rodriguez's analysis was that the most important variable in explaining the variation in the dependent variable (a dummy variable that measures whether the higher ranked player won the match) was the difference in official WTA or ATP rankings for the players in the match; this was the only significant variable across all models for both genders. They find that past success in the previous year's tournament has large explanatory power for men but not women. Additionally, they find that historical rankings (having previously been ranked in the top 10) are much more important explanatory factors for women than men. Physical characteristics such as age and height are significant across genders. Last, when computing forecasting accuracy of the model, the most important factor to include is past performance.

Our Data

To replicate the work of Corral and Prieto-Rodriguez we collected men's and women's tennis match data from Jeff Sackmann of Tennis Abstract. This provided us with game-level data on every tournament for both men and women from 2000-2015, including all Grand Slams. It also included official ranking data from the WTA and ATP. From this data on player rankings and match results, we were able to generate our dependent variable: HIGHERRANKEDVICTORY, a binary variable measuring if the higher ranked player won the match.

Next, we created dummy variables for each of the Grand Slam tournaments so we could look at Grand Slam-specific data and control for differences in the tournaments, especially their different surfaces. We also created dummy variables for each round of the tournament to test the finding that "the larger the difference in prizes between winner and loser, the more unlikely it becomes for a lower-ranked player to win" (Corral and Rodriguez, 553). The coefficients on these dummies are expected to increase as the rounds of the tournament move forward. In other words, the higher the stakes, the higher the expected probability that the higher ranked player will win.

We then created variables to control for past performances and differences in player's rankings as they entered the tournament (DIFRANKING). This was a logarithmic difference, to account for the changes in important for different levels of ranking. Since the Grand Slams are played on different surfaces (Australian and U.S. on hard courts, Wimbledon on grass, and French on clay) it is important to

control for those differences in the analysis. We also created the variable DIFROTOUR, which measures the differences in rounds achieved by the higher and lower ranked player in the previous year of that same tournament. This coefficient will be positive if player's perform better on some types of surfaces than others. To further control for past performances, we created the dummy variable EXTOP10H, which takes on a value of 1 if the higher ranked player was a former top-ten player; EXTOP10L does the same for the lower ranked player. Finally we control for player's physical characteristics of age, height and right vs. left-handedness.

We ended up with 2,032 observations for the 2005-2008 Grand Slams across genders; this is the data used to replicate Corral and Prieto-Rodriguez's findings. Our data also includes the years 2000-2015, which we will use to improve upon the original analysis. Additionally, we will use our data on wins, losses and game scores to generate the NCAA's Ratings Percentage Index, chess' Elo rating, and Google's PageRank for our data. We will then test each of these rankings performance in relation to the official ATP and WTA systems used by Corral and Prieto-Rodriguez.

Summary Statistics

Corral and Prieto-Rodriguez's ten most important statistics for both men's and women's grand slam data are summarized in table 1 below. They had 2,022 observations for men, 1,930 observations for women, and reported means and standard deviations for both sets. Our replication of the summary statistics for means and standard deviations are similar in both magnitude and sign to Corral and Prieto-Rodriguez's findings, as can be seen in table 2 and 3 for men and women respectively.

Table 1
Descriptive statistics.

	Male		Female	
	Mean	Std. dev.	Mean	Std. dev.
HIGHER-RANKED VICTORY	0.712	0.453	0.715	0.452
DIFRANKING	1.440	1.081	1.399	1.051
EXTOP10H	0.262	0.440	0.245	0.430
EXTOP10L	0.115	0.319	0.065	0.247
DIFHEIGHT (m.)	−0.002	0.091	0.025	0.091
DIFHEIGHT2 (m.)	0.008	0.014	0.009	0.012
DIFAGE	−0.112	4.897	0.070	5.767
DIFAGE2	23.978	34.341	33.246	45.016
LEFTL	0.110	0.313	0.077	0.267
LEFTH	0.094	0.293	0.064	0.244
BOTHLEFT	0.021	0.143	0.004	0.064
BOTHRIGHT	0.774	0.418	0.855	0.352
Number of observations	2022		1930	

Table 1

Variable	Obs	Mean	Std. Dev.	Min	Max
higherrank~y	2,032	.7145669	.4517316	0	1
difranking	2,032	1.444417	1.086722	.0129034	6.041207
extop10h	2,032	.4694882	.499191	0	1
extop10l	2,032	.113189	.3169016	0	1
difheight	2,032	-.0031102	.0898194	-.33	.35
difheight2	2,032	.0080732	.0131852	0	.1225
difage	2,032	-.1314276	4.921697	-16.09583	17.50856
difage2	2,032	24.22846	35.05803	.0000675	306.5496
leftl	2,032	.0748031	.2631382	0	1
lefth	2,032	.0260827	.1594204	0	1
bothleft	2,032	.0201772	.1406406	0	1
bothright	2,032	.7746063	.4179442	0	1

Table 2

An interesting finding from these statistics is that the higher ranked player wins their match roughly the same percentage of the time for both men and women (~71.5%). Corral and Prieto-Rodriguez found one significant difference between the summary statistics for men and women was the number of players who had been ranked in the top ten at some point in the previous 5 years; for men this number was roughly 29%, whereas for women it was only 15%. They attributed this difference to the fact that male tennis players generally have longer careers, so male players ranked in the top ten in the previous 5 years are more likely to still be active players (Corral and Prieto-Rodriguez 554).

Variable	Obs	Mean	Std. Dev.	Min	Max
higherrank~y	2,032	.7199803	.4491191	0	1
difranking	2,032	1.415864	1.066629	.0094787	5.545177
extop10h	2,032	.4330709	.4956222	0	1
extop10l	2,032	.0871063	.2820602	0	1
difheight	2,032	.0111417	.0699847	-.28	.3200001
difheight2	2,032	.0050196	.0098138	0	.1024
difage	2,032	.1547115	5.716089	-23.4579	16.33676
difage2	2,032	32.68153	45.09786	.0001199	550.2733
leftl	2,032	.0708661	.2566643	0	1
lefth	2,032	.0595472	.2367043	0	1
bothleft	2,032	.003937	.0626374	0	1
bothright	2,032	.8134843	.3896182	0	1

Table 3

We had slightly more observations than the paper which could have accounted for some slight variation in our summary statistics, but generally we believe our data matches well with the data that the

paper used. After running these summary statistics we felt confident that we had been able to replicate Corral and Prieto-Rodriguez's data and could confidently replicate their analysis.

Probit Regression Results

Corral and Prieto-Rodriguez's ran three separate probit regressions on the dependent variable HIGHERRANKEDVICTORY, found in Tables 4 and 5. The first model included all match and physical characteristics, the second excluded differences in past performances, and the third excluded physical characteristics. Since the authors used probit regression models, the magnitude of the coefficient is not very useful for interpreting its effect; instead, we looked to the sign of the coefficient to see the direction of the relationship and its significance level to judge its importance.

The most significant finding was the effect of the variable DIFRANKING on the dependent variable. The coefficient on this variable was the only one to be significant across models. The coefficient is positive for both men and women, meaning that the larger the difference in the ranking between the higher and lower player, the larger the probability that the higher ranked player will win. This is intuitive, when very high ranked players play very low ranked players there is a very small probability of an upset.

Probit results for women.

	M1			M2			M3		
	Coeff.	St. dev.	ME	Coeff.	St. dev.	ME	Coeff.	St. dev.	ME
DIFRANKING	0.384***	0.044	0.123				0.410***	0.041	0.132
EXTOP10H	0.453***	0.095	0.133				0.415***	0.090	0.124
EXTOP10L	−0.562***	0.146	−0.203				−0.498***	0.143	−0.179
DIFROTOUR	0.003	0.018	0.001				−0.006	0.017	−0.002
DIFHEIGHT	0.646	0.411	0.207	1.799***	0.385	0.604			
DIFHEIGHT2	2.463	3.269	0.790	4.661	3.087	1.564			
DIFAGE	−0.019***	0.006	−0.006	−0.003	0.006	−0.001			
DIFAGE2	0.000	0.001	0.000	−0.001	0.001	0.000			
LEFTL	0.023	0.121	0.007	0.023	0.117	0.008			
LEFTH	−0.074	0.126	−0.024	−0.182	0.122	−0.064			
BOTHLEFT	−0.450	0.463	−0.162	−0.172	0.450	−0.061			
2ND ROUND	−0.007	0.081	−0.002	0.152**	0.077	0.050	0.008	0.078	0.003
3RD ROUND	−0.172*	0.104	−0.057	0.045	0.096	0.015	−0.157	0.102	−0.052
4TH ROUND	−0.170	0.139	−0.057	0.030	0.127	0.010	−0.156	0.137	−0.052
QUARTERFINAL	−0.008	0.195	−0.003	0.060	0.176	0.020	0.010	0.193	0.003
SEMIFINAL	−0.743***	0.256	−0.278	−0.798***	0.229	−0.304	−0.724***	0.256	−0.270
FINAL	−0.544	0.350	−0.199	−0.495	0.318	−0.184	−0.547	0.347	−0.201
AUSTRALIA	−0.259***	0.092	−0.086	−0.251***	0.088	−0.087	−0.263***	0.089	−0.088
FRENCH OPEN	−0.150	0.091	−0.049	−0.136	0.088	−0.047	−0.150*	0.089	−0.050
WIMBLEDON	−0.266***	0.091	−0.089	−0.224**	0.088	−0.078	−0.253***	0.089	−0.085
CONSTANT	0.249***	0.089		0.652***	0.078		0.222***	0.078	
Number of observations		1930			1930			2032	
Likelihood ratio test		257			74			252	
Pseudo-R^2		0.111			0.031			0.104	
Log-likelihood		−1025			−1117			−1089	

Table 4

The effect of players having previously been ranked in the top ten does not appear to be as significant of a factor for men as it is for women. The coefficient EXTOP10H is highly significant and positive for women but not for men. EXTOP10L, if the lower ranked player was previously in the top ten, has a negative and significant coefficient for women in both models it is included in and only the first

model for men. This can be interpreted as the probability of an upset is larger if the lower ranked player was previously ranked in the top ten, we were surprised to see this not have a large effect for men.

For tournament-specific effects, DIFROTOUR is significant for men but insignificant for women. This reinforces the theory that men have more surface-biased skills than women. Controlling for previous round reached in the tournament is insignificant in all cases for men but significant and negative for semifinalists and 3rd round reachers for women.

Probit results for men.

	M1			M2			M3		
	Coeff.	St. dev.	ME	Coeff.	St. dev.	ME	Coeff.	St. dev.	ME
DIFRANKING	0.321***	0.039	0.104				0.342***	0.038	0.112
EXTOP10H	0.129	0.080	0.041				−0.005	0.075	−0.002
EXTOP10L	−0.373***	0.106	−0.130				−0.113	0.098	−0.038
DIFROTOUR	0.081***	0.017	0.026				0.071***	0.017	0.023
DIFHEIGHT	0.314	0.340	0.102	0.052	0.326	0.018			
DIFHEIGHT2	−2.364	2.189	−0.764	−3.955*	2.110	−1.338			
DIFAGE	−0.050***	0.007	−0.016	−0.037	0.006	−0.013			
DIFAGE2	0.002**	0.001	0.001	0.001	0.001	0.000			
LEFTL	−0.176*	0.100	−0.059	−0.093	0.095	−0.032			
LEFTH	−0.035	0.111	−0.012	−0.051	0.104	−0.017			
BOTHLEFT	0.023	0.228	0.007	0.058	0.214	0.019			
2ND ROUND	0.029	0.078	0.009	0.183**	0.074	0.060	0.044	0.077	0.014
3RD ROUND	−0.058	0.098	−0.019	0.039	0.093	0.013	−0.075	0.097	−0.025
4TH ROUND	−0.068	0.132	−0.022	0.064	0.125	0.021	−0.061	0.131	−0.020
QUARTERFINAL	0.118	0.198	0.037	0.290	0.185	0.090	0.128	0.194	0.040
SEMIFINAL	−0.124	0.253	−0.042	0.069	0.242	0.023	−0.156	0.246	−0.053
FINAL	0.048	0.375	0.015	0.061	0.340	0.020	−0.053	0.359	−0.018
AUSTRALIA	0.110	0.089	0.035	0.107	0.085	0.036	0.123	0.088	0.039
FRENCH OPEN	−0.043	0.087	−0.014	−0.049	0.084	−0.017	−0.042	0.086	−0.014
WIMBLEDON	−0.010	0.088	−0.003	−0.037	0.084	−0.013	0.003	0.086	0.001
CONSTANT	0.056	0.088		0.514***	0.075		0.027	0.078	
Number of observations		2022			2022			2032	
Likelihood ratio test		253			57			197	
Pseudo-R^2		0.104			0.023			0.081	
Log-likelihood		−1088			−1186			−1120	

Table 5

We were surprised to see most of the variables controlling for physical characteristics are insignificant in the models for women. In the first model only DIFAGE is negatively significant (younger players have an advantage), however this effect is lost in the second model. Only the coefficient on DIFHEIGHT is significant in this second model (taller players have a higher chance of an upset). For men, DIFAGE is negative and significant in the first model, as is LEFTL. This means the lower ranked players has a higher probability of an upset if he is younger and, interestingly, if he is left handed and playing against a right handed player.

Tables 6 and 7 show our regression results for women's and men's data respectively. We replicated the three separate probit models for both genders and found similar results to Corral and Prieto-Rodriguez. For women the most important variables were again DIFRANKING, EXTOP10H and EXTOP10L, which are highly significant in both models they are included in. DIFRANKING has a positive coefficient and is significant at p value<.001, EXTOP10H and EXTOP10L have positive and negative coefficients respectively and were again significant at p value<.001. These results are consistent

with the findings of Corral and Prieto-Rodriguez. Our regression findings support the author's theory that men have more surface-biased skills because the coefficient on DIFROTOUR is significant for men and insignificant for women.

	(1) higherr~y		(2) higherr~y		(3) higherr~y	
higherra~y						
difranking	0.354***	(7.77)			0.374***	(8.23)
extop10h	0.446***	(4.72)			0.434***	(4.65)
extop10l	-0.484***	(-3.42)			-0.447**	(-3.20)
difrotour	-0.00536	(-0.30)			-0.0165	(-0.94)
difheight	0.688	(1.43)	1.590***	(3.49)		
difheight2	4.079	(1.16)	4.179	(1.24)		
difage	-0.0180**	(-3.14)	-0.00987	(-1.86)		
difage2	-0.000189	(-0.27)	-0.000193	(-0.29)		
leftl	0.00213	(0.02)	0.0339	(0.29)		
lefth	-0.127	(-1.00)	-0.221	(-1.81)		
bothleft	-0.397	(-0.86)	-0.172	(-0.38)		
round2	-0.0122	(-0.15)	0.185*	(2.49)	-0.00854	(-0.11)
round3	-0.284**	(-2.65)	0.0250	(0.27)	-0.268*	(-2.52)
round4	-0.244	(-1.63)	0.0586	(0.46)	-0.242	(-1.62)
quarterf~l	-0.0910	(-0.43)	0.0613	(0.35)	-0.0839	(-0.40)
semifinal	-0.724**	(-2.69)	-0.717**	(-3.14)	-0.722**	(-2.68)
final	-0.500	(-1.42)	-0.537	(-1.69)	-0.513	(-1.46)
australia	-0.264**	(-2.95)	-0.259**	(-3.03)	-0.273**	(-3.06)
french	-0.127	(-1.41)	-0.141	(-1.64)	-0.135	(-1.51)
wimbledon	-0.257**	(-2.86)	-0.228**	(-2.66)	-0.255**	(-2.86)
_cons	0.260**	(3.02)	0.695***	(9.27)	0.256**	(3.24)
N	2032		2032		2032	
R-sq						
adj. R-sq						

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Table 6

Our physical characteristic regression results also matched the original paper's findings, DIFHEIGHT was positive and significant at p <.001 in the second model and DIFAGE was negative and significant at p<.01 in the first model. Similarly, for men's data there is a negative coefficient on DIFAGE. There is a slight deviation in our findings for physical characteristics from Corral and Prieto-Rodriguez's findings in that LEFTH and LEFTL were omitted from the model because STATA reported that these variables perfectly predicted the change in the dependent variable. We think this strange finding may be due to our men's data lacking complete physical characteristic data for many observations. We did not run into this problem with the women's data.

Generally, the sign and statistical significance of the coefficients are quite similar between our regression results and Corral and Prieto-Rodriguez's for both men and women.

	(1) higherr~y		(2) higherr~y		(3) higherr~y	
higherra~y						
difranking	0.298***	(6.74)			0.321***	(7.75)
extop10h	0.0567	(0.68)			0.0645	(0.82)
extop10l	-0.0564	(-0.49)			0.00716	(0.07)
difrotour	0.0915***	(5.02)			0.0754***	(4.44)
difheight	0.552	(1.52)	0.168	(0.48)		
difheight2	-3.246	(-1.36)	-4.939*	(-2.14)		
difage	-0.0406***	(-5.60)	-0.0361***	(-5.36)		
difage2	0.00205*	(2.07)	0.00167	(1.74)		
leftl	.	.	.	.		
lefth	.	.	.	.		
bothleft	0.0460	(0.20)	0.0468	(0.22)		
round2	0.0427	(0.52)	0.198*	(2.53)	0.0489	(0.63)
round3	-0.128	(-1.21)	-0.00197	(-0.02)	-0.111	(-1.11)
round4	-0.0587	(-0.41)	0.102	(0.78)	-0.113	(-0.82)
quarterf~l	0.237	(1.07)	0.425*	(2.11)	0.0640	(0.32)
semifinal	0.141	(0.49)	0.330	(1.24)	-0.130	(-0.49)
final	-0.196	(-0.48)	-0.110	(-0.31)	-0.130	(-0.35)
australia	0.104	(1.12)	0.0863	(0.96)	0.127	(1.45)
french	-0.0366	(-0.40)	-0.0534	(-0.60)	-0.0419	(-0.49)
wimbledon	-0.0122	(-0.13)	-0.0253	(-0.28)	0.00802	(0.09)
_cons	0.0194	(0.21)	0.493***	(6.40)	0.0190	(0.24)
N	1827		1827		2032	
R-sq						
adj. R-sq						

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Table 7

In-Sample Predictive Accuracy

Corral and Prieto-Rodriguez looked to the predictive accuracy of their models by testing the predicted values against in-sample data and then the forecasting capability of the models by extending the findings to out of sample time periods(tables 8 and 9). Generally, the findings from this section show that their models do a good job at predicting wins, but the best models are those that include a player's past performance.

The in-sample predictive accuracy for men (2005–2008).

Predicted interval	M1		M2		M3	
	Actual outcome		Actual outcome		Actual outcome	
	Und. victory	Fav. victory	Und. victory	Fav. victory	Und. victory	Fav. victory
0–0.5	**94**	98	**2**	3	**46**	42
	(49%)	(51%)	**(40%)**	(60%)	**(52%)**	(48%)
0.5–1	489	**1341**	581	**1436**	538	**1406**
	(27%)	**(73%)**	(29%)	**(71%)**	(28%)	**(72%)**
0.2–0.3	**1**	0				
	(100%)	(0%)				
0.3–0.4	**18**	16			**2**	2
	(53%)	(47%)			**(50%)**	(50%)
0.4–0.5	**75**	82	**2**	3	**44**	40
	(48%)	(52%)	**(40%)**	(60%)	**(52%)**	(48%)
0.5–0.6	156	**181**	47	**62**	167	**210**
	(46%)	**(54%)**	(43%)	**(57%)**	(44%)	**(56%)**
0.6–0.7	154	**262**	273	**534**	189	**337**
	(37%)	**(63%)**	(34%)	**(66%)**	(36%)	**(64%)**
0.7–0.8	111	**316**	226	**619**	117	**366**
	(26%)	**(74%)**	(27%)	**(73%)**	(24%)	**(76%)**
0.8–0.9	55	**325**	34	**214**	50	**298**
	(14%)	**(86%)**	(14%)	**(86%)**	(14%)	**(86%)**
0.9–1	13	**257**	1	**7**	15	**195**
	(5%)	**(95%)**	(13%)	**(88%)**	(7%)	**(93%)**
Pearson χ^2	219			51		169
Brier score	0.183			0.199		0.187

Table 8

An important statistic in these tables is the Brier score. This statistic is an indicator of predictive accuracy and is defined as the sum of the difference between the predicted probability that the higher-ranked player wins and the dummy variable if the higher ranked player wins divided by total number of observations. The higher this number is, the worse job the model did at predicting the outcome of games. The Brier score for men and women was highest in the second model, when past performances were excluded from the model. The first and third model's Brier score were very similar for both men and women and therefore the variables measuring past performance seem to be the most important when looking at predictive accuracy.

The in-sample predictive accuracy for women (2005–2008).

Predicted interval	F1		F2		F3	
	Actual outcome		Actual outcome		Actual outcome	
	Und. victory	Fav. victory	Und. victory	Fav. victory	Und. victory	Fav. victory
0–0.5	**91**	57	**22**	16	**72**	46
	(61%)	(39%)	**(58%)**	(42%)	**(61%)**	(39%)
0.5–1	459	**1323**	528	**1364**	508	**1406**
	(26%)	**(74%)**	(28%)	**(72%)**	(27%)	**(73%)**
0–0.1	**1**	0				
	(100%)	(0%)				
0.1–0.2	**5**	0			**7**	0
	(100%)	(0%)			**(100%)**	(0%)
0.2–0.3	**13**	4	**1**	1	**9**	2
	(76%)	(24%)	**(50%)**	(50%)	**(82%)**	(18%)
0.3–0.4	**13**	10	**8**	6	**14**	11
	(57%)	(43%)	**(57%)**	(43%)	**(56%)**	(44%)
0.4–0.5	**59**	43	**13**	9	**42**	33
	(58%)	(42%)	**(59%)**	(41%)	**(56%)**	(44%)
0.5–0.6	133	**189**	37	**52**	162	**201**
	(41%)	**(59%)**	(42%)	**(58%)**	(45%)	**(55%)**
0.6–0.7	164	**267**	236	**457**	178	**316**
	(38%)	**(62%)**	(34%)	**(66%)**	(36%)	**(64%)**
0.7–0.8	99	**305**	207	**598**	100	**328**
	(25%)	**(76%)**	(26%)	**(74%)**	(23%)	**(77%)**
0.8–0.9	44	**282**	48	**237**	50	**299**
	(14%)	**(87%)**	(17%)	**(83%)**	(14%)	**(86%)**
0.9–1	19	**280**	0	**20**	18	**262**
	(6%)	**(94%)**	(0%)	**(100%)**	(6%)	**(94%)**
Pearson χ^2	243			64		237
Brier score	0.178			0.196		0.180

Table 9

Table 10 and 11 show our replication of these findings by Corral and Prieto-Rodriguez for men and women respectively. Our findings are generally consistent with those of the original paper; the most important factor in predictive accuracy is past performance in both sets of models.

Our chi squared and Brier score results were both consistent with the authors findings, being best for the first and third models and worst for the second model for both men and women, reinforcing that the second model does the worst job of the three for predicting results.

Men

	model 1		model 2		model 3	
	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory
P(higher rank wins)						
0-10						
20-30	**3**	0				
	(10.00%)	(0.00%)				
30-40	**10**	11			**3**	0
	(47.62%)	(52.38%)			**(10.00%)**	(0.00%)
40-50	**62**	59	**3**	1	**39**	40
	(51.24%)	(48.76%)	**(75.00%)**	(25.00%)	**(49.37%)**	(50.63%)
50-60	137	**182**	42	**58**	170	**203**
	(42.95%)	**(57.05%)**	(42.00%)	**(58.00%)**	(45.58%)	**(54.42%)**
60-70	148	**242**	256	**504**	181	**336**
	(37.95%)	**(62.05%)**	(33.68%)	**(66.32%)**	(35.01%)	**(64.99%)**
70-80	110	**293**	191	**533**	122	**360**
	(27.30%)	**(72.70%)**	(26.38%)	**(73.62%)**	(25.31%)	**(74.69%)**
80-90	49	**285**	34	**190**	50	**310**
	(14.67%)	**(85.33%)**	(15.18%)	**(84.82%)**	(13.89%)	**(86.11%)**
90-100	8	**228**	1	**14**	15	**203**
	(33.90%)	**(96.61%)**	(66.70%)	**(93.33%)**	(68.80%)	**(93.12%)**
	Pearson chi2= 194.8994		Pearson chi2 = 47.3741		Pearson chi2 = 178.4929	
	Brier score: 0.2186		Brier Score: 0.2359		Brier Score: 0.2133	

Table 10

Women

	model 1		model 2		model 3	
	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory
P(higher rank wins)						
0-10	**1**	0			**1**	0
	(100.00%)	(0.00%)			**(100.00%)**	(0.00%)
20-30	**7**	0			**8**	0
	(100.00%)	(0.00%)			**(100.00%)**	(0.00%)
30-40	**15**	13	**16**	12	**11**	9
	(53.57%)	(46.43%)	**(57.14%)**	(42.86%)	**(55.00%)**	(45.00%)
40-50	**69**	54	**16**	12	**43**	38
	(56.10%)	(43.90%)	**(57.14%)**	(42.86%)	**(53.09%)**	(46.91%)
50-60	154	**181**	29	**39**	170	**206**
	(45.97%)	**(54.03%)**	(42.65%)	**(57.35%)**	(45.21%)	**(54.79%)**
60-70	156	**301**	219	**438**	173	**320**
	(34.14%)	**(65.86%)**	(33.33%)	**(66.67%)**	(35.09%)	**(64.91%)**
70-80	99	**300**	249	**755**	94	**304**
	(24.81%)	**(75.19%)**	(24.80%)	**(75.20%)**	(23.62%)	**(76.38%)**
80-90	48	**284**	52	**208**	49	**260**
	(14.46%)	**(85.54%)**	(20.00%)	**(80.00%)**	(15.86%)	**(84.14%)**
90-100	20	**330**	0	**8**	20	**326**
	(5.71%)	**(94.29%)**	(0.00%)	**(100.00%)**	(5.78%)	**(94.22%)**
	Pearson chi2= 258.4414		Pearson chi2 = 47.7113		Pearson chi2 = 234.3296	
	Brier score:0.2154		Brier Score: 0.2408		Brier Score: 0.2160	

Table 11

Out of Sample Forecasting Accuracy

The models' forecasting accuracies are displayed in tables 12 and 13. The authors train their models using 2005-2008 data, and then evaluate their performanced based on unseen data from the 2009 Australian Open. As with the in-sample data, the second models (that exclude past performance) are the worst forecasters of results for both men and women. This can be seen through the highest Brier score and lowest chi squared score of the three models. The best model for forecasting both men and women's game outcomes is the first model, controlling for all factors. Further, the authors found the physical characteristic variables are more important to include in the model for men than for women.

The out-of-sample forecasting accuracy for men (2009 Australian Open).

Predicted interval	M1		M2		M3	
	Actual outcome		Actual outcome		Actual outcome	
	Und. victory	Fav. victory	Und. victory	Fav. victory	Und. victory	Fav. victory
0.2–0.3	**0**	1				
	(0%)	(100%)				
0.3–0.4	**2**	1				
	(67%)	(33%)				
0.4–0.5	**6**	1			**2**	1
	(86%)	(14%)			**(67%)**	(33%)
0.5–0.6	6	**7**	1	**1**	5	**8**
	(46%)	**(54%)**	(50%)	**(50%)**	(38%)	**(62%)**
0.6–0.7	6	**19**	10	**14**	14	**21**
	(24%)	**(76%)**	(42%)	**(58%)**	(40%)	**(60%)**
0.7–0.8	5	**14**	14	**47**	8	**22**
	(26%)	**(74%)**	(23%)	**(77%)**	(27%)	**(73%)**
0.8–0.9	4	**28**	4	**25**	4	**24**
	(13%)	**(88%)**	(14%)	**(86%)**	(14%)	**(86%)**
0.9–1	1	**17**	1	**1**	0	**18**
	(6%)	**(94%)**	(50%)	**(50%)**	(0%)	**(100%)**
Pearson χ^2	26		7		15	
Brier score	0.158		0.182		0.167	

Table 12

The out-of-sample forecasting accuracy for women (2009 Australian Open).

Predicted interval	F1		F2		F3	
	Actual outcome		Actual outcome		Actual outcome	
	Und. victory	Fav. victory	Und. victory	Fav. victory	Und. victory	Fav. victory
0.1–0.2	**0**	1			**0**	1
	(0%)	(100%)			**(0%)**	(100%)
0.2–0.3	**0**	2			**0**	2
	(0%)	(100%)			**(0%)**	(100%)
0.3–0.4	**1**	1	**0**	1	**0**	1
	(50%)	(50%)	**(0%)**	(100%)	**(0%)**	(100%)
0.4–0.5	**4**	5	**0**	2	**2**	3
	(44%)	(56%)	**(0%)**	(100%)	**(40%)**	(60%)
0.5–0.6	8	**11**	3	**3**	14	**16**
	(42%)	**(58%)**	(50%)	**(50%)**	(47%)	**(53%)**
0.6–0.7	7	**16**	18	**39**	11	**18**
	(30%)	**(70%)**	(32%)	**(68%)**	(38%)	**(62%)**
0.7–0.8	8	**6**	9	**27**	4	**17**
	(57%)	**(43%)**	(25%)	**(75%)**	(19%)	**(81%)**
0.8–0.9	2	**19**	1	**3**	4	**20**
	(10%)	**(90%)**	(25%)	**(75%)**	(17%)	**(83%)**
0.9–1	2	**14**	1	**0**	0	**14**
	(13%)	**(88%)**	(100%)	**(0%)**	(0%)	**(100%)**
Pearson χ^2	15		5		16	
Brier score	0.175		0.219		0.193	

Table 13

Our replication of this analysis can be found in tables 14 and 15 for men and women respectively. Our results were consistent with the findings of Corral and Prieto-Rodriguez, with the worst model being the second and the best was the first (which was very similar to the third). This reinforces their idea that the most important factor for forecasting accuracy is past performances for men and women in the data.

Out-Of-Sample Men (2009 Australian Open)

	model 1		model 2		model 3	
	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory
P(higher rank wins)						
0-10						
20-30						
30-40	**1**	1				
	(50.00%)	(50.00%)				
40-50	**4**	1				
	(80.00%)	(20.00%)				
50-60	6	**3**	2	**1**	6	**7**
	(66.67%)	**(33.33%)**	(66.67%)	**(33.33%)**	(46.15%)	**(53.85%)**
60-70	8	**19**	13	**15**	13	**21**
	(29.63%)	**(70.37%)**	(46.43%)	**(53.57%)**	(38.24%)	**(61.67%)**
70-80	7	**20**	12	**43**	9	**23**
	(25.93%)	**(74.07%)**	(21.82%)	**(78.18%)**	(28.13%)	**(71.88%)**
80-90	5	**18**	4	**2**	5	**24**
	(21.74%)	**(78.26%)**	(15.38%)	**(84.62%)**	(17.24%)	**(82.76%)**
90-100	1	**20**	1	**1**	0	**19**
	(4.76%)	**(95.24%)**	(50.00%)	**(50.00%)**	(0.00%)	**(100.00%)**
	Pearson chi2= 19.9957		Pearson chi2 = 10.5006		Pearson chi2 = 13.3022	
	Brier score: 0.1720		Brier Score: 0.1934		Brier Score: 0.1671	

Table 14

Out-Of-Sample Women (2009 Australian Open)

	model 1		model 2		model 3	
	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory
P(higher rank wins)						
0-10						
20-30	**0** **(0.00%)**	3 (100.00%)			**0** **(0.00%)**	2 (100.00%)
30-40	**1** **(25.00%)**	3 (75.00%)	**0** **(0.00%)**	1 (100.00%)	**0** **(0.00%)**	3 (100.00%)
40-50	**5** **(50.00%)**	5 (50.00%)	**0** **(0.00%)**	3 (100.00%)	**3** **(50.00%)**	3 (50.00%)
50-60	10 (41.67%)	**14** **(58.33%)**	3 (50.00%)	**3** **(50.00%)**	12 (40.00%)	**18** **(60.00%)**
60-70	6 (28.57%)	**15** **(71.43%)**	17 (25.76%)	**49** **(74.24%)**	9 (33.33%)	**18** **(66.67%)**
70-80	6 (22.22%)	**21** **(77.78%)**	10 (25.64%)	**29** **(74.36%)**	4 (18.18%)	**18** **(81.82%)**
80-90	2 (10.53%)	**17** **(89.47%)**	2 (16.67%)	**10** **(83.33%)**	4 (18.18%)	**18** **(81.82%)**
90-100	2 (10.53%)	**17** **(89.47%)**			0 (0.00%)	**15** **(100.00%)**
	Pearson chi2= 12.3214 Brier score: 0.1896		Pearson chi2 = 3.7841 Brier Score: 0.1969		Pearson chi2 = 14.2803 Brier Score: 0.1861	

Table 15

Analysis Improvements

To improve upon the analysis, we first extended Corral and Prieto-Rodriguez's work to include more years; their study only looked at 2005-2008, and we have collected data from 2000 to 2015. We re-ran the experiment using 2000-2014 as in-sample, and 2015 as out-of-sample data. Table 16 shows the probit regression results for men's data for years 2000-2014. Generally, significance levels improved for variables that were statistically significant in the previous models, which reinforces our original findings. One interesting finding from including more data in the model is that the variables EXTOP10H and EXTOP10L have both become highly significant across models. In our original model, neither of these variables were significant, which we felt was a strange result. We expected previous top ten experience would give the player a significantly higher chance of victory, as was the result in the women's data. The model that includes more years intuitively makes sense and shows a more complete picture of the relationship between these factors and match outcomes. Another result that changed with additional data was that the variable measuring the difference in height became highly significant across models, another result that makes intuitive sense.

	(1) higherr~y		(2) higherr~y		(3) higherr~y	
higherra~y						
difranking	0.308***	(13.66)			0.325***	(15.31)
extop10h	0.125**	(2.85)			0.128**	(3.12)
extop10l	-0.261***	(-4.63)			-0.228***	(-4.31)
difrotour	0.0717***	(7.67)			0.0588***	(6.73)
difheight	0.584**	(3.21)	0.597***	(3.39)		
difheight2	-2.809*	(-2.29)	-3.695**	(-3.11)		
difage	-0.0250***	(-6.90)	-0.0171***	(-5.05)		
difage2	-0.000222	(-0.42)	-0.000130	(-0.26)		
leftl	.	.	.	.		
lefth	.	.	.	.		
bothleft	-0.0447	(-0.37)	-0.0419	(-0.36)		
round2	-0.000825	(-0.02)	0.131**	(3.27)	0.00775	(0.20)
round3	-0.0422	(-0.76)	0.0883	(1.72)	-0.0280	(-0.53)
round4	-0.0142	(-0.19)	0.116	(1.70)	-0.0270	(-0.37)
quarterf~l	0.0303	(0.29)	0.152	(1.62)	-0.0292	(-0.29)
semifinal	0.0443	(0.32)	0.0630	(0.50)	-0.0420	(-0.32)
final	-0.0668	(-0.35)	-0.118	(-0.66)	-0.0835	(-0.46)
australia	0.0609	(1.27)	0.0418	(0.91)	0.0631	(1.40)
french	-0.0118	(-0.25)	-0.0173	(-0.38)	-0.0108	(-0.24)
wimbledon	-0.0776	(-1.63)	-0.0876	(-1.92)	-0.0749	(-1.68)
_cons	0.102*	(2.21)	0.552***	(13.84)	0.0697	(1.73)
N	6745		6745		7620	

Table 16

We also reran the probit regressions for women's data with the years 2000-2014, which can be found below in table 17. Interestingly, unlike with the men's data, the women's results did not change in a significant way by including more years. The main change that came from including more data was that the coefficients on significant variables became even more significant, supporting our original findings.

	(1) higherr~y		(2) higherr~y		(3) higherr~y	
higherra~y						
difranking	0.312***	(14.77)			0.320***	(15.18)
extop10h	0.278***	(6.28)			0.256***	(5.85)
extop10l	-0.427***	(-6.24)			-0.371***	(-5.53)
difrotour	0.0168	(1.90)			0.0113	(1.29)
difheight	0.246	(0.96)	0.987***	(4.00)		
difheight2	2.511	(1.38)	2.157	(1.22)		
difage	-0.0161***	(-5.28)	-0.00629*	(-2.20)		
difage2	-0.000348	(-0.98)	-0.0000971	(-0.28)		
leftl	-0.0567	(-0.97)	-0.0190	(-0.33)		
lefth	-0.0649	(-1.07)	-0.123*	(-2.10)		
bothleft	-0.265	(-1.38)	-0.253	(-1.36)		
round2	-0.0380	(-0.96)	0.117**	(3.12)	-0.0354	(-0.90)
round3	-0.0846	(-1.59)	0.121*	(2.47)	-0.0686	(-1.29)
round4	-0.0848	(-1.15)	0.0813	(1.25)	-0.0771	(-1.05)
quarterf~l	0.0376	(0.36)	0.109	(1.22)	0.0374	(0.36)
semifinal	-0.311*	(-2.33)	-0.287*	(-2.43)	-0.301*	(-2.27)
final	-0.303	(-1.71)	-0.395*	(-2.42)	-0.314	(-1.76)
australia	-0.0950*	(-2.11)	-0.103*	(-2.38)	-0.0925*	(-2.06)
french	-0.125**	(-2.77)	-0.122**	(-2.82)	-0.122**	(-2.72)
wimbledon	-0.118**	(-2.63)	-0.120**	(-2.76)	-0.110*	(-2.45)
_cons	0.212***	(4.99)	0.607***	(16.27)	0.186***	(4.68)
N	7620		7620		7620	

Table 17

The in-sample prediction results for this expanded range are shown in tables 18 and 19 respectively. The results were similar to the analysis of the original years. The highest Brier score for both genders is found in model 2 where past performance metrics are excluded. Therefore, we reaffirm again that these variables are important factors in predictive accuracy.

A difference we found in this analysis is that the difference in the Brier score across models has gone down relative to the original analysis. This means that the models have converged by adding more data and their predictive capabilities have become more similar than they were originally.

	Men					
	model 1		model 2		model 3	
	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory
p(higher ranked victory)	0	1	0	1	0	1
20-30	**2**	1				
	66.67%	33.33%				
30-40	**32**	18			**13**	7
	64.00%	36.00%			**65%**	35%
40-50	**198**	168	**4**	3	**132**	108
	54.10%	45.90%	**57.14%**	42.86%	**55%**	45%
50-60	522	**663**	62	**80**	642	**808**
	44.05%	**55.95%**	43.66%	**56.34%**	44.28%	**55.72%**
60-70	594	**1116**	815	**1612**	694	**1285**
	34.74%	**65.26%**	33.58%	**66.42%**	35.07%	**64.93%**
70-80	375	**1116**	1,071	**3048**	410	**1317**
	25.15%	**74.85%**	26%	**74%**	23.74%	**76.26%**
80-90	197	**990**	7	**43**	228	**1144**
	16.60%	**83.40%**	14%	**86%**	16.62%	**83.38%**
90-100	39	**714**			53	**779**
	5.18%	**94.82%**			6.37%	**93.63%**
	chi2 = 607.8764	Brier = 0.2145	chi2 = 65.6358	Brier = .2337	chi2 = 628.9296	Brier = 0.212

Table 18

	Women					
	model 1		model 2		model 3	
	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory
p(higher ranked victory)	0	1	0	1	0	1
20-30	**1**	0			1	0
	100%	0.00%			100%	0%
30-40	**22**	13			**20**	7
	62.86%	37.13%			**74%**	26%
40-50	**122**	118	**0**	3	**64**	69
	50.83%	49.17%	**0.00%**	100.00%	**48%**	52%
50-60	662	**794**	63	**76**	732	**887**
	45.47%	**54.53%**	45.32%	**54.68%**	45.21%	**54.79%**
60-70	757	**13931**	964	**1926**	733	**1397**
	35.21%	**64.79%**	33.36%	**66.64%**	34.41%	**65.59%**
70-80	370	**1197**	1,136	**3259**	388	**1162**
	23.61%	**76.39%**	26%	**74%**	25.03%	**74.97%**
80-90	203	**1147**	31	**162**	204	**1198**
	15.04%	**84.96%**	16%	**84%**	14.55%	**85.46%**
90-100	57	**764**			52	**706**
	6.94%	**93.06%**			6.86%	**93.14%**
	chi2 = 656.0552	Brier = 0.2123	chi2 = 82.948	Brier = .2362	chi2 = 626.6802	Brier = .2124

Table 19

To test whether adding more data to the models improved their forecasting accuracy we generated Brier and chi squared scores for their predicted outcomes relative to actual outcomes in the 2015 Australian Open across genders; these findings can be found in tables 20 and 21. Again, we found very similar results to the original analysis; adding more years to the data reinforced our original findings that past performances are the most important variable in forecasting accuracy. Adding more years did not necessarily increase the predictive capabilities of the models, the results are essentially the same.

	Out-Of-Sample Women (2015 Australian Open)					
	Model 1		Model 2		Model 3	
	Underdog Victory	Favorite Vicrtory	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory
p(higher ranked wins)	0	1	0	1	0	1
30-40	**1**	0				
	100%	0%				
40-50	**6**	1			**3**	2 5
	85.71%	14.29%			**60%**	40%
50-60	5	**18**	0	**3**	12	**16**
	21.74%	**78.26%**	0%	**100%**	42.86%	**57.14%**
60-70	17	**22**	24	**39**	11	**2**
	43.59%	**56.41%**	38.10%	**61.90%**	31.43%	**68.57%**
70-80	8	**16**	14	**45**	9	**17**
	33.33%	**66.67%**	23.73%	**76.27%**	34.62%	**65.38%**
80-90	1	**21**	1	**1**	3	**18**
	4.55%	**95.45%**	50%	**50%**	14.29%	**85.71%**
90-100	1	**10**			1	**11**
	9.09%	**90.91%**			8.33%	**91.67%**
	chi2 = 25.6914	Brier = 0.1924	chi2 = 4.6456	Brier 0.2106	chi2= 9.6384	Brier = 0.1963

Table 20

	Out-Of-Sample Men (2015 Australian Open)					
	model 1		model 2		model 3	
	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory	Underdog Victory	Favorite Victory
p(higher ranked victory)	0	1	0	1	0	1
40-50	**4**	3			0	**2**
	57.14%	42.86%			0.00%	**100%**
50-60	5	**14**			7	**14**
	26.32%	**73.68%**			33.33%	**66.67%**
60-70	9	**17**	9	**26**	15	**18**
	34.62%	**65.38%**	25.71%	**74.29%**	45.45%	**54.55%**
70-80	7	**23**	19	**55**	7	**24**
	23.33%	**76.67%**	25.68%	**74.32%**	22.58%	**77.42%**
80-90	2	**9**	0	**3**	2	**20**
	18.18%	**81.82%**	0%	**100%**	9.09%	**90.91%**
90-100	1	**18**			1	**17**
	5.26%	**94.74%**			5.56%	**94.44%**
	chi2 = 9.4213	Brier = .1756	chi2 = 1.0275	Brier = .1863	chi = 15.4209	Brier = 1709

Table 21

Finally, we experimented with three additional ratings systems to see if any of them showed promise for outperforming the official ATP ratings in predictive ability. These three ratings were the RPI, Elo, and PageRank. RPI is a rating system used in the NCAA equal to .25(win%) + .5(opponents win%) + .25(opponents opponents' win%). Elo is a system typically used in chess ratings where a player's rating improves or decreases depending on the corresponding rating of the player he faces. This makes Elo an interesting rating system for tennis, since it factors in strength of schedule implicitly by basing the value of a match upon the assigned value of the opponent. Finally, PageRank is an algorithm used in graph theory, designed by Google to rank the relative importances of web pages on the internet. PageRank, in this case, represents a tennis season as a directed graph of players and their opponents, resulting in a weight that corresponds to a player's "importance" in the graph, which can in turn be interpreted as a rating of player ability.

We ran simple probit models of our new rankings regressed on new dependent variables (each system's ratings difference) to see which models performed best. While RPI performed poorly, Elo and PageRank exceeded expectations, each achieving higher pseudo-R-squared values than the official ATP rankings in explaining Grand Slam match victors.

	(1) favwins~r		(2) favwins~o		(3) favwins~i		(4) higherr~y	
main								
PageRank	245.6***	(13.85)						
Elo			0.00770***	(13.97)				
RPI					1.344***	(5.85)		
difranking							0.465***	(12.68)
_cons	-0.0549	(-1.00)	0.0545	(0.96)	0.484***	(11.09)	-0.0153	(-0.29)
N	1905		1905		1708		1905	
pseudo R~q	0.103		0.140		0.019		0.086	

Table 22

While this is only a single-variable regression, we believe that these high scores for PageRank and Elo show promising potential for use as tennis ratings systems in the future, seeing that they each beat the official ATP rating benchmark by a considerable margin.

Conclusion

In conclusion, through the collection and manipulation of Jeff Sackmann's tennis match data across both genders from 2005 and 2015, we were able to accurately replicate the statistical analysis found in *Are differences in ranks good predictors for Grand Slam tennis matches?* And to produce very similar results. Our statistics across means, standard deviations, and coefficients matched those found by Corral and Prieto-Rodriguez in both magnitude and sign.

Both our results, and those of Corral and Prieto-Rodriguez indicate that differences in ranks are the crucial element to explaining variation in the dependent variable; the higher ranked player winning the match. Another important finding is that player's past performances are the most important variables to include in models for predicting real game outcomes.

Given funding and an extended time period, the analysis could be improved through a variety of initiatives, including adding more observations, looking at tournaments other than Grand Slams, and filling in the many missing values in the women's physical characteristics data. Additionally, given the promise that both Elo and PageRank scores have shown in preliminary analysis, we would examine these systems in greater depth, improving them to take final scores of matches into account.

References

Dyte, D., & Clarke, S. (2000). A ratings based Poisson model for World Cup soccer simulation. *The Journal of the Operational Research Society*, *51*(8), 993–998.

Elo, A. E. (2008). *The rating of chess players, past and present.* San Rafael, United States: Ishi Press.

Leitner, C., Zeileis, A., & Hornik, K. (2008). *Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008.* Report 26, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. URL: http://epub.wu.ac.at/.

Leitner, C, Zeileis, A., & Hornik, K. (2008). *Who is going to win the EURO 2008? (A statisitical investigation of bookmakers odds).* Report 65, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series. URL: http://epub.wu.ac.at/.