Assessing Human Skill in Video Soccer

Joe Miller
jollierme@gmail.com
Updated 4/9/2010

A crude method of assessing skill and producing handicaps in video soccer competitions is presented. Definitions of player rating, relative skill, and handicap rating are given. Data and resulting calculations from two competitions are provided.

In *Sensible World of Soccer*[1] (SWOS), each team has a different overall level of ability. It is based on the computer-defined abilities of all the individuals on a team. Ability has numerous components, such as speed, ball control, and finishing ability. The overall ability of a team is expressed in Stars.[2] The wide range of Stars of the teams in SWOS provides a way to handicap human versus human matches and tournaments. A player with little skill can play with a team that has a large number of Stars to make matches more competitive.

Table 1 contains the results of SWOS II[3] and each of the quantities defined below for all participants.[4] The Number of Points (Pts) a player earns in a tournament is three for a Win (W), one for a Draw (D), and zero for a Loss (L).[5] Points Per Game (PPG) is defined as the quotient of Pts and the number of Matches (Games) played. In 39 of 43 decisive matches played, the player with the eventual higher PPG won the match.

In determining a player's rating, Strength of Opponents (SO) can be considered. SO is the average PPG of all of a player's opponents, excluding the match(es) that the player and each of his opponents played against each other.[6] The quotient of PPG and "the maximum PPG across all players" is the Normalized PPG (Norm PPG). Normalized SO (Norm SO) is calculated in the same way. PPG and SO are assumed to be equally important in determining a rating,[7] so calculating the Norm values is necessary.[8] The average of Norm PPG and Norm SO is the Tournament Rating (Rating).

---

[1] The XBOX 360 version (a remake of the *Sensible World of Soccer '96/'97* video game).
[2] xflea's (*SWOS 96/97 Database* v.1.0 - http://xflea.republika.pl/swos/swos-database.zip) calculated values from the *stars* column in the *teams* sheet are used in this discussion.
[3] SWOS II was an all-human tournament played on March 7, 2009.
[4] Throughout this discussion, it is assumed that the distributions of various quantities are normal. That is not necessarily the case. No effort was made to correct non-normalities.
[5] While this scheme is good for encouraging attacking soccer, it may not be the best basis for calculating skill level or handicaps.
[6] This avoids double-counting match results, since the results of such matches are already included in a player's PPG.
[7] This assumption is arbitrary. One might argue that PPG or SO should have a bigger influence on rating than the other. One might also argue that the variation in the population distributions should be incorporated into the calculation of Rating.
[8] This is because the average PPG and the average SO are not equal - all players did not play the same number of matches, and the SO calculation excludes certain matches.

In 41 of 43 decisive matches played, the player with the eventual higher Rating won the match. This would seem to indicate that Rating is a better winner predictor than PPG alone. However, more data is needed to validate this conclusion. Because of the nature of the calculations,[9] Ratings are only valid for the tournament results from which they are calculated. However, Rating can be used to approximate the relative skill of each player and to determine the teams that players should use to make future matches more competitive.

Though Rating indicates how well a player performed in a tournament, it does not indicate their skill (how well they play SWOS) compared to others.[10] A number of factors could be used along with Rating to determine skill. The only two factors that are readily quantifiable are Stars and the total number of SWOS matches played. Since the total number of matches played is less than twenty for all players, it is assumed that the number of matches played is much less relevant than Stars.

A player's Skill is the quotient of Rating and Stars. Using Microsoft Excel, a linear least-squares regression was completed to determine the relationship between Rating and Stars (see Figure 1). The result indicates an increase in Rating as Stars increases. The R^2 value of the regression (about 0.2) is relatively low. Removing two potential outliers improves the R^2 value to about 0.5 but tightens the confidence intervals little. Outliers are not excluded because there is no clear reason to do so.

The regression residual[11] can be used to show how well a player did relative to other players in one tournament, considering the teams played. Relative Skill is the quotient of the residual and "the difference between the individual upper 95% Rating and the regression-predicted Rating."[12,13]

Skill calculated from one tournament might be used to level the playing field in future tournaments. One method assigns teams to players by equalizing each player's Handicap Rating, which is the product of Skill (from a previous tournament) and Stars (of the team played with in the next competition). An alternative method assigns teams so that players with more skill have a slightly higher Handicap Rating than opponents with less skill.

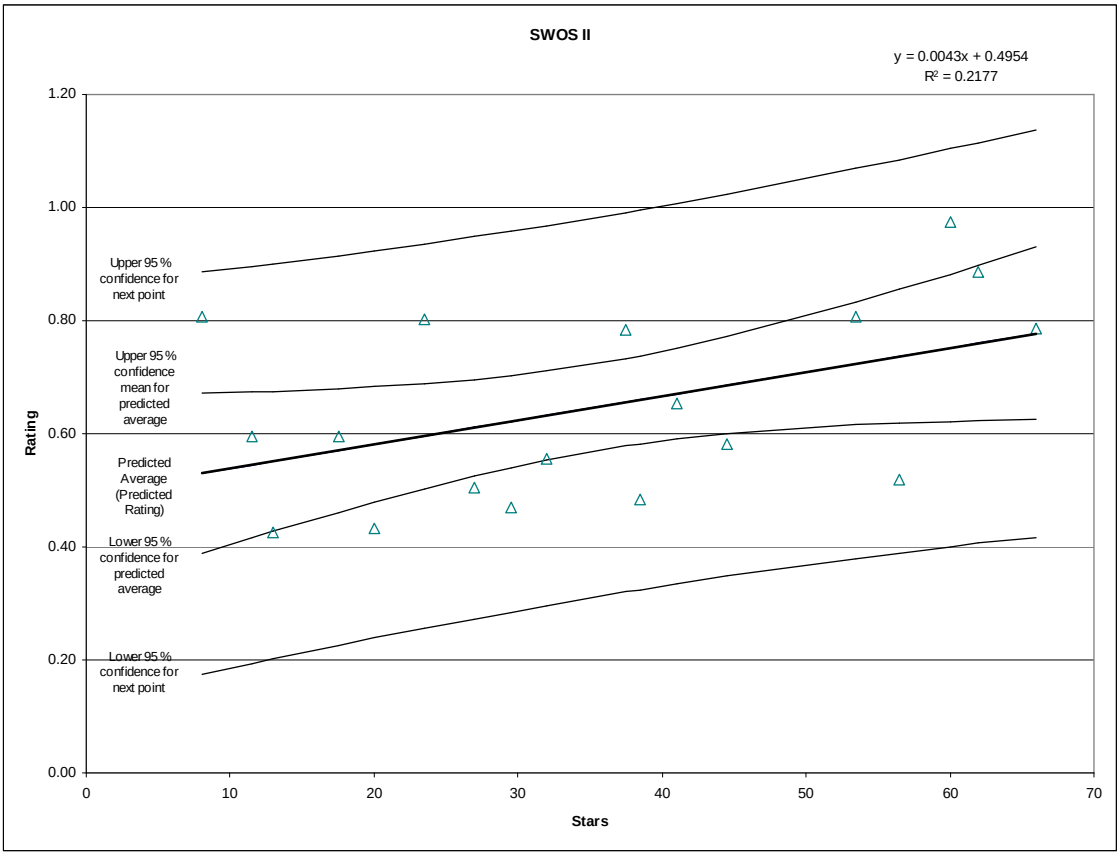
[9] Variation in Stars across players and the number of draws in a tournament (due to the 3/1/0 point system) make direct comparisons of ratings in one tournament to another difficult.
[10] If all teams have the same Stars, then Rating and Skill are the same.
[11] Also called "error", which is the difference between the actual Rating and the regression-predicted Rating at a given Stars value.
[12] This accounts for the edge effects of the 95% intervals, which get wider at the edges of the Stars domain.
[13] One problem with this approach is that the maximum rating a player can achieve is limited by their SO. Another problem is that the individual upper 95% rating is above 1 for some players, and a rating above 1 cannot be achieved.

Table 1. SWOS II Results and Calculations (each line is for a different player):

W	L	D	Pts	Games	PPG	SO	Norm PPG	Norm SO	Rating	Stars	Skill	Relative Skill
5	1	1	16	7	2.29	1.32	0.83	0.78	0.81	8.0	0.101	0.78
7	0	1	22	8	2.75	1.60	1.00	0.95	0.97	60.0	0.016	0.63
3	2	1	10	6	1.67	1.68	0.61	1.00	0.80	23.5	0.034	0.61
3	1	2	11	6	1.83	1.51	0.67	0.90	0.78	37.5	0.021	0.38
6	2	0	18	8	2.25	1.61	0.82	0.96	0.89	62.0	0.014	0.36
5	1	1	16	7	2.29	1.32	0.83	0.78	0.81	53.5	0.015	0.24
2	4	0	6	6	1.00	1.39	0.36	0.83	0.60	11.5	0.052	0.14
1	2	2	5	5	1.00	1.39	0.36	0.83	0.60	17.5	0.034	0.07
4	2	0	12	6	2.00	1.42	0.73	0.84	0.79	66.0	0.012	0.02
2	3	1	7	6	1.17	1.48	0.42	0.88	0.65	41.0	0.016	-0.05
0	2	3	3	5	0.60	1.50	0.22	0.89	0.56	32.0	0.017	-0.23
1	4	1	4	6	0.67	1.55	0.24	0.92	0.58	44.5	0.013	-0.31
1	3	1	4	5	0.80	1.21	0.29	0.72	0.50	27.0	0.019	-0.31
0	4	1	1	5	0.20	1.31	0.07	0.78	0.43	13.0	0.033	-0.36
0	3	2	2	5	0.40	1.21	0.15	0.72	0.43	20.0	0.022	-0.43
1	3	1	4	5	0.80	1.09	0.29	0.65	0.47	29.5	0.016	-0.45
1	2	2	5	5	1.00	1.01	0.36	0.60	0.48	38.5	0.013	-0.53
1	4	0	3	5	0.60	1.38	0.22	0.82	0.52	56.5	0.009	-0.63

Pts = 3 * W + D
PPG = Pts / Games
SO = Average PPG of opponents after removing the results of game(s) played against them
Norm Quantity = Quantity / Maximum Quantity of all players
Rating = (Norm PPG + Norm SO) / 2
Skill = Rating / Stars
Relative Skill = Regression Residual / (Individual Upper 95% Rating - Predicted Rating)

Figure 1. Rating vs. Stars Regression



Addendum 1

SWOS III was an all-human tournament played on January 23, 2010. In 50 of 56 decisive matches played, the player with the eventual higher PPG won the match. In 52 of 56 decisive matches played, the player with the eventual higher Rating won the match. When two players faced each other twice, they were treated as new opponents in the SO calculation.

Table 2. SWOS III Results and Calculations:

W	L	D	Pts	Games	PPG	SO	Norm PPG	Norm SO	Rating	Stars	Skill	Relative Skill
3	2	3	12	8	1.50	1.90	0.63	1.00	0.81	31.0	0.026	0.78
4	1	1	13	6	2.17	1.35	0.90	0.71	0.81	36.0	0.022	0.68
6	1	1	19	8	2.38	1.83	0.99	0.96	0.98	66.0	0.015	0.66
2	1	2	8	5	1.60	1.37	0.67	0.72	0.69	20.5	0.034	0.57
6	1	2	20	9	2.22	1.60	0.93	0.84	0.88	61.5	0.014	0.47
2	3	1	7	6	1.17	1.48	0.49	0.78	0.63	20.0	0.032	0.40
2	2	2	8	6	1.33	1.66	0.56	0.87	0.71	44.0	0.016	0.25
2	0	2	8	4	2.00	1.51	0.83	0.80	0.81	64.0	0.013	0.22
2	1	1	7	4	1.75	1.38	0.73	0.73	0.73	49.0	0.015	0.21
4	1	0	12	5	2.40	1.09	1.00	0.57	0.79	60.5	0.013	0.19
5	1	1	16	7	2.29	1.27	0.95	0.67	0.81	65.0	0.012	0.19
3	1	2	11	6	1.83	1.44	0.76	0.76	0.76	64.0	0.012	0.06
1	4	0	3	5	0.60	1.54	0.25	0.81	0.53	23.5	0.023	0.04
2	3	1	7	6	1.17	1.10	0.49	0.58	0.53	25.0	0.021	0.02
1	1	2	5	4	1.25	1.46	0.52	0.77	0.64	48.5	0.013	-0.04
2	2	2	8	6	1.33	1.36	0.56	0.72	0.64	48.5	0.013	-0.06
2	3	1	7	6	1.17	0.97	0.49	0.51	0.50	24.0	0.021	-0.07
1	3	1	4	5	0.80	1.49	0.33	0.79	0.56	39.0	0.014	-0.14
2	3	1	7	6	1.17	1.25	0.49	0.66	0.57	45.0	0.013	-0.20
0	2	2	2	4	0.50	1.65	0.21	0.87	0.54	39.0	0.014	-0.20
2	1	2	8	5	1.60	0.97	0.67	0.51	0.59	49.5	0.012	-0.23
0	4	1	1	5	0.20	1.46	0.08	0.77	0.43	30.0	0.014	-0.39
0	3	2	2	5	0.40	1.10	0.17	0.58	0.37	26.0	0.014	-0.48
0	3	2	2	5	0.40	1.16	0.17	0.61	0.39	29.5	0.013	-0.50
2	3	0	6	5	1.20	1.01	0.50	0.53	0.52	58.0	0.009	-0.59
0	3	2	2	5	0.40	1.07	0.17	0.56	0.36	36.0	0.010	-0.69
0	3	1	1	4	0.25	0.99	0.10	0.52	0.31	57.5	0.005	-1.20

Figure 2. Rating vs. Stars Regression