University of the Free State

Department of Mathematical Statistics and Actuarial Science

Introducing Momentum to the Elo rating System

D.W. Bester and M.J. von Maltitz

Keywords: Elo, relative rating system, chess rating, simulation

Abstract

Arpad E. Elo (1903-1992) developed the Elo rating system in the late 1950's to rate chess players' performance. It is still widely used today for the rating of chess players. This paper aims to investigate a possible improvement of the Elo system, namely the introduction of momentum into the rating system. It is of interest to establish whether this can be done while harnessing the simplicity of the basic Elo system. Simplicity is as important as effectiveness if one wishes to expand the reach of the Elo rating system. The system holds great potential for measuring two-player games, as well as team games, and is not restricted to win, draw or loss outcomes, but can be used for ratings in systems with score-based games as well. With the current immense internet gaming arena kept in mind, there truly exist opportunities for dynamic (yet simple) rating systems to flourish.

Contents

Introduction

The Elo rating system, which is still in use today, was developed by Arpad E. Elo (1903-1992) in the late 1950's. His initial research took place whilst heading a committee that studied the then used Harkness System with the aim to improve it. This led to Elo developing a formula that emulated the results of the Harkness System, thus creating the Elo system. Subsequently, the United States Chess Federation (USCF) approved of Elo's new system in August 1960 and it was implemented (Sloan, 2008). Glickman (1995: 1) stated that, "The introduction of chess rating systems may have done more to popularize tournament chess than any other single factor." Since implementation of the Elo system it has been scrutinised on a regular basis, leading to the identification of areas for improvement. As a result of this, many modifications have been made to the Elo system (Glickman, 1995). Modifications are usually aimed at two aspects; *firstly*, it is important that a rating system succeeds in assigning a player an accurate rating within a small timeframe. Accuracy is considered as successfully measuring a player's relative strength (within the set of players in the system). Time will usually be measured by the number of games needed to achieve an accurate rating. *Secondly*, once an accurate rating is achieved, it is important to keep the deviation of this rating around the true underlying skill level to a minimum.

This paper aims to investigate a possible improvement of the Elo rating system, namely the introduction of momentum into this system. It is of interest to establish whether this can be done while retaining the simplicity of the basic Elo rating procedure. One can assume that chess players are familiar with Elo-type rating systems and therefore accept it. However, it may prove troublesome to introduce such systems to other gaming or sports environments, as players might be sceptical when it comes to scoring systems they are unfamiliar with. Thus, simplicity is as important as effectiveness if one wishes to expand the reach of the Elo rating system. Simplicity also ensures transparency. Thus, having a more effective, yet still simple rating system can be of great value, not only in the realm of tournament chess, but also, for example, in tennis. The latter has, similar to chess, a large number of individual players, who can also be ranked using a simple rating system. With the current immense internet gaming arena kept in mind, there truly exist opportunities for dynamic (yet simple) rating systems to flourish.

The objective of the study in this paper is to answer the following research questions:

- Will the addition of a momentum component make the Elo system more effective?
- Will the Elo system still be relatively simple with the addition of a momentum component?

The next section includes a literature review on the origins of the Elo system, with the aim to introduce the basic structure of the system. It also mentions some of the modifications that have been made to the Elo system, as well as a few other systems used for rating players. The subsequent chapter will elaborate on the methodology used to answer the research questions posed, including the methods used for simulating data testing this data. A chapter on experimentation, simulation and analysis will follow, where various methods of incorporating momentum will be applied and tested. Finally, a conclusion will end the paper, discussing the results obtained.

Literature review

Since the early days of competitive chess, the need for a dynamic rating system was present. Kenneth Harkness designed the Harkness system in the late 1940's, the starting point for all chess rating systems used today (Sloan, 2008). His system was very crude and had many shortcomings, but Arpad Elo used players' Harkness ratings to formulate the assumptions underlying the Elo rating system. The Harkness system can be summed up in a single table:

Rating Difference	Higher rated wins: Add to winner and deduct from loser	Lower rated wins: Add to winner and deduct from loser	Draw: Add to lower rated and deduct from higher rated
0 to 24	16	16	0
25 to 49	15	17	1
50 to 74	14	18	2
75 to 99	13	19	3
100 to 124	12	20	4
125 to 149	11	21	5
150 to 174	10	22	6
175 to 199	9	23	7
200 to 224	8	24	8
225 to 249	7	25	9
250 to 274	6	26	10
275 to 299	5	27	11
300 or more	4	28	12

Source: Ross (2007)

The crudeness of the system is immediately apparent, but it nonetheless supplied a rich database of players' ratings for Arpad Elo to analyse.

Elo (1978: 19) assumed that the performances of individual chess players were Normally distributed, and he did extensive studies to validate this (Elo, 1965 and McClintok, 1977, cited in Elo, 1978: 19). Elo used a mean of zero and a standard deviation, σ, of one class, or 200 points, for the distribution of individual player performances. For a game with two players the standard deviation of the

difference of performance would be $\sqrt{\sigma^2 + \sigma^2} = \sigma\sqrt{2} = 282.84$. Elo could use the difference between the rating of a player and the rating of the competition to calculate the expected outcome of any given game or games. The competition could be a single player or a collection of opponents, where the latter uses the average rating of the opponents in calculations. Elo (1978) also stated that the Logistic distribution could also be used as underlying model for individual performance. Today, the USCF uses the Logistic distribution, whilst FIDE (*Fédération Internationale des Échecs* or World Chess Federation) still uses the Normal distribution that Elo originally based his system on. The USCF uses the Logistic distribution as they regard it to most accurately extrapolate outcomes (Ross, 2007). The research proposed in this paper will make use of the Logistic curve as mentioned in Glickman and Jones (1999: 2).

Being able to calculate the expected outcome of any given game, Elo devised a range of formulas to calculate the changes in ratings. These formulae include methods for periodic measurements, continuous measurements, linear approximation formulae and rating of round robins (Elo, 1978: 24-29). Continuous measurement is of interest, as it allows for the most recent rating. The formula states:

$$R_n = R_o + k(W - W_e)$$

where R_n is the new rating and R_o is the pre-event rating. The variable W is the actual event outcome, 1 for each game won, 0 for each game lost and 0.5 for a game drawn.[1] The expected outcome of the event is W_e. The k-factor controls the magnitude of the rating change for the current event.[2] A constant k-factor for all events weighs all events as equally important. It is common practice to give high k-factors to new players and then to lower the factor as the number of games played increases, the reasoning being that after many games a player's listed rating will be close to the player's true underlying rating. When the listed rating nears the real rating, rating adjustments after each event need not be of a considerable magnitude. An event can represent a single game or a number of games. However, adjusting ratings after every game will provide more accurate results.

Adaptations of Elo's system used today concentrate on the underlying distribution of performance and various combinations of k-factors. Various techniques of assigning initial ratings are also explored. Players each have a rating and the number of games played should also be recorded, as it

[1] Even though the system is not limited to win/draw/loss outcomes, but can also handle scored outcomes, this paper is not considering this possibility. As mentioned in the conclusion, this is a topic that could be handled in further research.

[2] It is easy to see that the maximum rating change that can occur in a game is limited to the value of k, which is the amount of rating change when a win occurs for the player who has an infinitesimally small expectation of winning.

is used in determining k-factors. An example of a widely used and effective adaptation is the Glicko system, as stipulated in Glickman (2001). This system assigns each player a rating and a rating deviation. The rating deviation decreases with number of games played and is increased by time passing without playing rated games. Whilst this is a very effective system, calculations can be quite complex, especially without the aid of a computer. It is interesting to note that the Elo system is a special case of the Glicko system.

This system, along with the original Elo system, is described by Coulom (2008: 1) as an Incremental Rating System.[3]

A summary of the different rating systems in use, as described by Coulom (2008: 1-2), is as follows:

- Static Rating Systems, not considering the variation in time of players' ratings.
- Incremental rating systems, storing a small amount of data for each player used to calculate rating systems. Examples are the Elo, Glicko and TrueSkill systems.
- Decayed History Rating Systems. These give decaying weights to results of older games, deriving the current rating by assuming the most recent games hold more up to date information.
- Accurate Bayesian Inference, a model similar to incremental algorithms, but with less approximations. Coulom (2008) classifies the Whole History Rating system described in his paper as such.

These systems have been scrutinised and compared with each other, each having unique strengths and weaknesses. Though all of them are by all means effective, the mathematics behind calculating and estimating ratings can become very complex.

It is important to note that all these systems use relative ratings. The true rating of a player can never be known, so ratings are adjusted relative to each other. Two groups of players who compete independently – *i.e.* none of the players from one group play against players in the other group – will produce ratings which could not be compared between groups. For the purposes of this paper, a two-player world will be considered and the ratings of the players will vary relative to each other.

There remains a need to test the Elo system with the addition of a momentum component assigned alongside the ratings. The main focus should be on effectiveness accompanied by simplicity if the

[3]Coulom (2008) also classifies the Trueskill rating system as described by Herbrich and Graepel (2006) in this category.

system is to be introduced as a chess rating system as well as to sports and gaming communities outside of chess.

Methodology

The research proposed in this paper will be an empirical study. All the data to be tested will be simulated using computer packages. Since all data are simulated, the control over the data will be high. Simulations carried out will provide numerical data containing information about game outcomes. Outcomes can then be rated using different rating systems, allowing for comparisons among the rating systems.

The questions to be answered in this research are concerned with which rating system performs best under given scenarios. Moreover, the questions will be answered empirically, analysing the created data and testing the relevant hypothesis.

The software package that will be used to simulate the data is Matlab (MATLAB, 2007). Initially, a list of players will be created, all players being assigned a true underlying rating. Game outcomes will be generated from these true ratings. Each player in the list will also be assigned a rating to be altered by the rating system, as well as statistic to track the number of games played. At this point, it is important to elaborate on the generation of game outcomes. This will be done by assuming that a player's performance follows a logistic distribution with the expected score for Player A, from Glickman and Jones (1999: 2), calculated as follows:

$$E_A = \frac{1}{1 + 10^{-(R_A - R_B)/400}}$$

In this equation, R_A is the true rating of Player A and R_B is the true rating of Player B. Let W_A be the event where Player A wins the match:

$$W_A \sim Bernoulli(E_A)$$

This means that once the expected score for Player A is determined, a random number generator on the $(0,1)$ interval is constructed to determine the game's outcome – if a random draw is less than the expected score for Player A, we score a win for Player A, and a loss for Player B. The reverse is true for a random draw greater than the expected score for Player A; in this case a loss is notated for Player A and a win for Player B. Note that the generator will only generate a win or loss; draws will not be considered. It is worth mentioning that the expected scores utilise the same distribution that will be used as the underlying distribution to calculate the rating changes by the Elo system. In other words, Elo assumed that this is the distribution that adequately describes game outcomes given

ratings. By using this distribution to model the game outcomes in the experiment we are in effect controlling variables that may cause an effective system to not perform well.

The effectiveness of a rating system will be tested in two ways for every proposed method of including momentum, namely speed of the rating system and stability of the ratings under the rating system. This will be done by first creating a universe with only two players. The two players will each be assigned true underlying ratings as well as 'published' ratings to be changed under the system in question. The latter will be referred to as the quoted rating. The starting point for the quoted ratings will be exactly in the middle of the true ratings. For example, for two players with real ratings 1300 and 1700, the starting point will be 1500 for both players. It is important to note that if the starting position for the assigned ratings is not exactly in the middle of the true ratings for both players, we cannot expect the assigned ratings to converge to the real ratings. This is because the Elo formula only takes into account the difference between to players' ratings and not the actual magnitude of each player's rating. This makes it difficult to simulate a tournament with more than two players with the objective of checking to see when assigned ratings reach true ratings.

Games will be simulated between the two players and ratings adjusted accordingly. Each time the number of games taken to achieve the true rating difference will be recorded. This will be the measure of the system's speed. The above process will be repeated for various rating differences and k-factor combinations, to test the system under a wide array of circumstances. Each time the procedure will be repeated $10\ 000$ times, after which the average number of games taken to reach the true ratings will be calculated. Then we will test the following hypothesis:

> $H_{0,1}$: The Elo system and Momentum adjusted system are equally effective, *i.e.* the average number of games taken to reach the true ratings is the same for both systems.

This will be done using a t-test of equal means. To do this it will be required to first do an F-test of equal variances, to determine which assumptions underlie the t-test. All significance tests will be done at a 95% confidence level. It is important to note that a t-test carried out on a very large sample will sometimes inaccurately reject the H_0 because of the way the test procedure is structured. Thus, the sample of 10 000 values will be used to calculate the averages, but the t-test will be carried out on only 100 values, selected at random.

Stability will be tested by starting the quoted ratings equal to the actual ratings and then letting the players play $10\ 000$ games. After each game the ratings will be adjusted and recorded, so that the overall average and standard deviation of the player's rating can be calculated. If the system is

stable, we would expect the average quoted rating to remain close to the real rating, with a small standard deviation. The average and standard deviation of the quoted ratings for the proposed momentum adjusted systems can be compared with that of the original Elo.

The above tests will be repeated for the following player pairs:

Payer A		Player B	Difference
1450	vs	1550	100
1400	vs	1600	200
1300	vs	1700	400
1200	vs	1800	600
1100	vs	1900	800
1000	vs	2000	1000

We perform these tests for k-factors of 10, 15, 25, 16, 24 and 32. The first three are used in the FIDE system whereas the last three are used by many National Federations. These are the two most commonly used k-factor combinations (Vovk, 2008). This will allow us to compare the speed and stability of various systems with the basic Elo and each other. However, each comparison of a proposed system will begin with a plot of the rating movements of the system, along with the rating movements of the same scenario rated using Elo's system. It should be possible to deduce whether the system harbours potential by first looking at the plot. If deemed to have potential, further testing can be done on the system.

In a two-player world it will be impossible to get the assigned ratings close to the true ratings if the conditions aren't controlled to achieve this, as the ratings assigned to a player are relative to the ratings of the other players. However, if the conditions aren't controlled, it would be possible to calculate the difference between the true real ratings for both players and compare this with the difference of the assigned ratings.

In real world situation a player's true rating will never be known, but in the theoretical situation of this paper it is important to control the variables that affect the assigned ratings' movements relative to the true ratings.

After testing the systems for speed and stability, another indication of effectiveness will be to see if a system can overcome some weakness of the Elo system. In this case, the system will have to have a speed and stability close to that of the Elo system, whilst having some characteristic that improves on it. Stability will be tested more thoroughly by calculating the root mean squared error (RMSE) for

both systems *i.e.* the root of the mean squared difference between the true rating and the assigned rating. The RMSEs for the Elo system can then be compared with that of the proposed system.

To calculate RMSE values, simulate a random value between 100 and $1\,000$ to use as difference between true ratings. Then simulate $1\,000$ games, starting the players' assigned ratings at the level equal to their true ratings, and then calculate the RMSE for both the Elo and the other system in question:

$$RMSE = \sqrt{\frac{1}{1\,000}\sum_{i=1}^{1\,000}(actual\ rating - assigned\ rating)^2}$$

Then, for comparison, calculate the difference between the RMSE's of the two systems:

$$RMSE_{System\ in\ question} - RMSE_{Elo}$$

This test is done $1\,000$ times for each player pair, for $1\,000$ different random player pairs, giving $1\,000\,000$ values for the difference in RMSE values. A positive difference indicates a higher RMSE for the other system, indicating that it is less stable than the Elo system.

Three systems will be tested in this paper. The *first* is called the Buffer system and it gradually builds up momentum for a player throughout play, which is then used to offset rating changes at a later stage. *Second*, the Switching Momentum system, assigns a player a momentum of -1, 1 or 0, for a losing streak, a winning streak and broken streak respectively. k-factor values are then doubled on winning or losing streaks. *Thirdly*, the Deficit system is tested; it tracks the movements of the Elo system, but creates a deficit for a player when a streak is broken. The player's rating will remain unchanged until the Deficit is recovered.

Experimentation, simulation and analysis

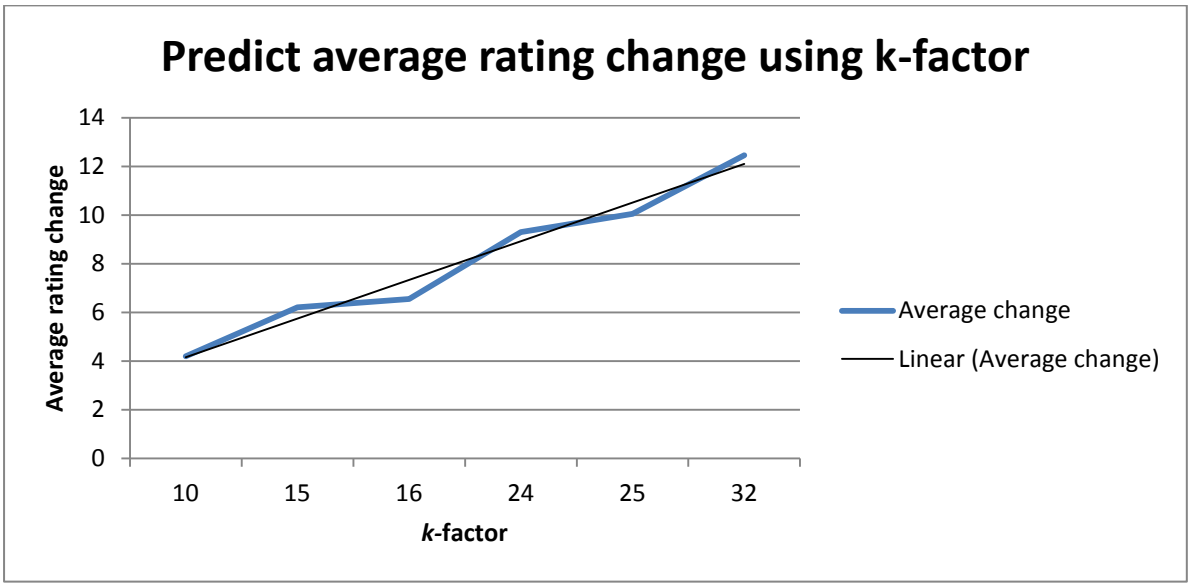
In order to explore different ways of implementing momentum, it is of interest to investigate how the chosen k-factor will influence the average rating changes experienced by the system. In other words, how much a player's rating will change on average after one game. This information can then be used in a system where a player builds momentum with each game, using it to protect him from a future loss. Knowing the magnitude of rating changes under specified k-factors is important when deciding where to stop the accumulation of momentum to ensure that it only protects a player against a small number of adverse rating changes. To do this, create eight players with real ratings as given in the table below.

Player	True Rating
1	2000
2	1950
3	1900
4	1850
5	1800
6	1750
7	1700
8	1650
9	1600
10	1550

Then give them each an assigned rating of 1200 and simulate 100 round robins, recording the rating change after every game, *i.e.* $k(W - W_e)$ in the Elo rating adjustment formula. Do this at the k-factors mentioned above. Using the results from the round robins, the following table of average rating changes is produced:

k-factor	Average change
10	4.187317707
15	6.206991263
16	6.54197403
24	9.30081315
25	10.05383182
32	12.45037077

Attempt to use a simple linear regression to predict the average rating changes:



Regression Statistics	
Multiple R	0.999

R Square	0.998
Adjusted R Square	0.998
Standard Error	0.147
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	45.416	45.416	2114.632	1.33E-06
Residual	4	0.086	0.021		
Total	5	45.501			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.526	0.176	2.996	0.045
k-factor	0.374	0.008	45.985	1.33E-06

From the output it is clear that the k-factor can successfully be used to predict the average rating change, as 99.8% of the variation in the average rating changes can be explained by the variation in the k-factorss. Thus the following equation will be adequate for predictions:

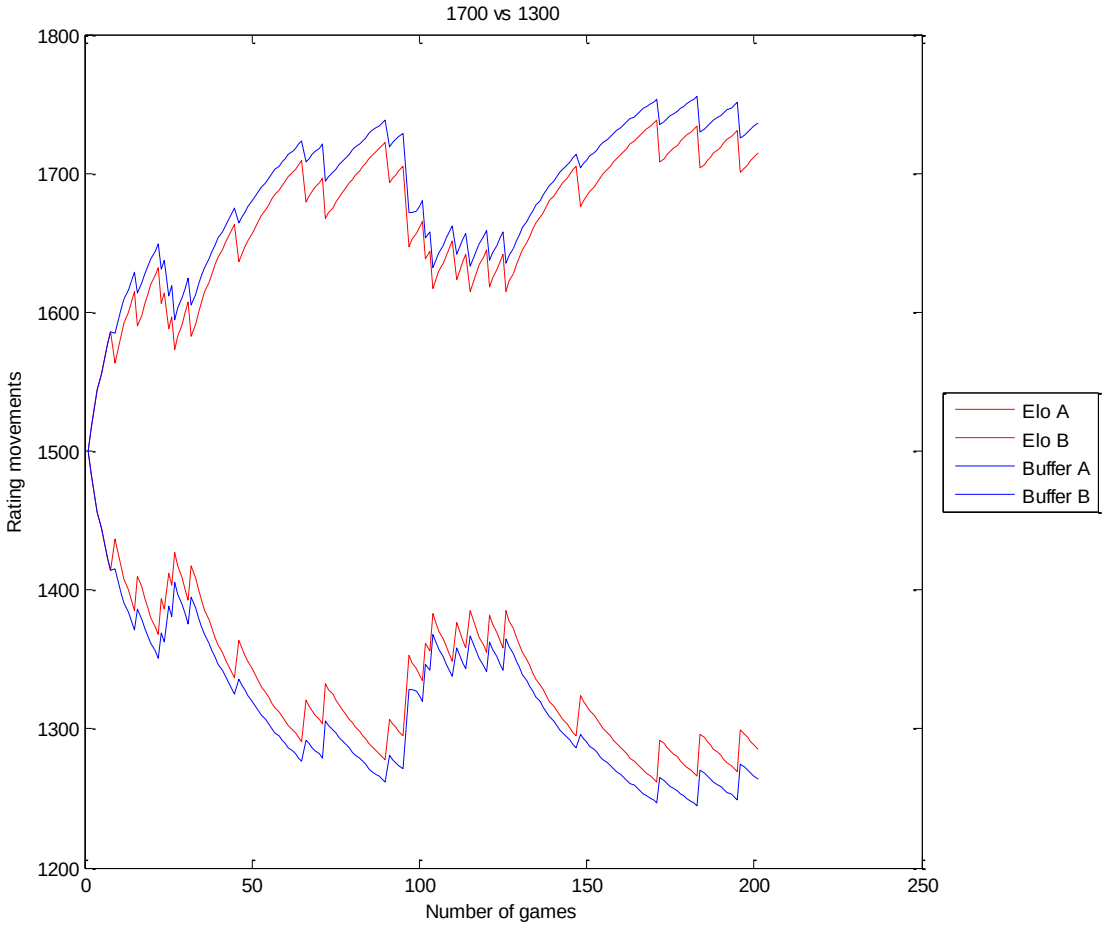
$$average\ rating\ change = 0.52644155 + 0.373628274 * k$$

The Buffer system

We begin by testing a simple and intuitive system, where a player builds momentum as he/she plays, then uses the momentum to protect him- or herself against a sudden rating change. Positive and negative momentum positions are possible. After each game, a player will gain momentum equal to a quarter of the rating change experienced in the game. It can also be altered to add other coefficients of the rating change to the momentum, but for the purposes of proof of concept in this paper, a value of ¼ will be used:

$$new\ momentum = old\ momentum + \frac{k}{4}(score - \ expected\ score)$$

This will accumulate up to a maximum momentum holding, calculated using the above formula. Thus, a player can at any time only hold enough momentum to buffer an average rating change. This can be altered; the goal is to see if such a system can be effective. When a player has positive momentum and loses a game, the rating change is applied to the momentum first. A rating change larger than the momentum at the current position will be carried over to the actual rating of the player, resulting in the momentum being reset to zero. In effect, players use momentum as a "shield" or "buffer" to protect them against large future rating movements. Test this system for two players with true ratings 1700 and 1300, firstly for speed of convergence. Both players start with an assigned rating of 1500, the following plot illustrates the rating movements.



The speed at which the system achieves the true ratings appears to be slightly greater than the basic Elo system, so it is worthwhile to test this. We simulate games under both systems to calculate the average number of games taken to reach the true rating difference. Do this for all the above mentioned player and k-value combinations. Each time, do a t-test of equal means to test the following hypothesis:

$$H_0: \mu_{Elo} = \mu_{Buffer}$$

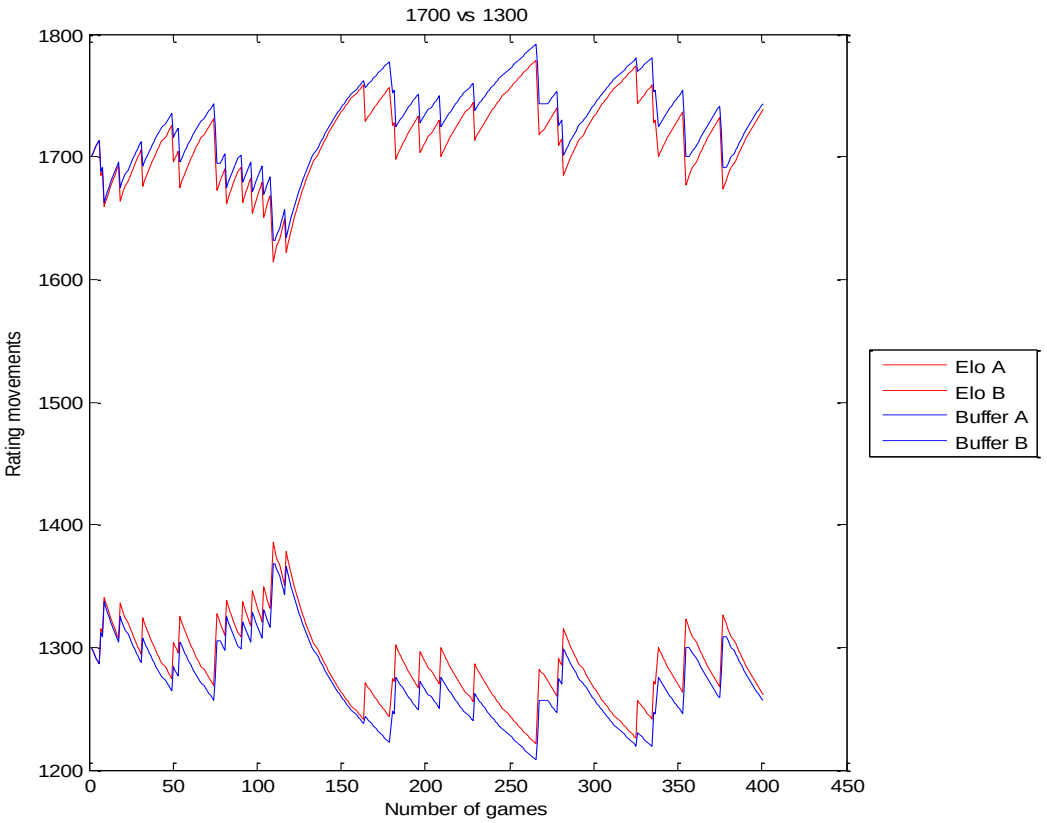
Where μ_i is the average number of games taken for the assigned ratings to reach the true ratings under each system.

	k = 10		k = 15		k = 16		k = 24		k = 25		k = 32	
Rating Diff	Elo	Buffer	Elo	Buffer	Elo	Buffer	Elo	Buffer	Elo	Buffer	Elo	Buffer
100												
Mean	62	46	37	29	34	27	19	17*	18	16*	13	12*
SD	40	29	26	20*	24	19*	15	12*	14	11	10	9
200												
Mean	100	70	61	45	57	42	35	27	33	26	24	20
SD	48	31	31	21	29	20	19	14	18	14	14	11
400												

Mean	244	168	150	108	139	102	85	65	81	62	60	47
SD	98	53	62	37	58	35	37	24	36	23	27	18
600												
Mean	671	462	415	300	379	279	236	179	222	172	168	131
SD	262	130	168	93	151	86	99	60	91	57	72	44
800												
Mean	1999	1376	1233	898	1146	834	694	534	667	514	495	390
SD	788	379	494	274	465	255	283	169	276	164	207	129
1000												
Mean	6211	4266	3814	2765	3565	2586	2168	1658	2069	1588	1524	1204
SD	2461	1192	1518	836	1462	783	900	540	868	525	634	396

To determine whether to use the assumption of equal variances, F-tests were performed in each case, and, every time the hypothesis of equal variances was rejected at a confidence level of 95%., a t-test assuming unequal variances was used for that pair. An asterisk (*) indicates an insignificant difference. Excluding the cases with small true rating differences and high k-values, the Buffer system outperformed the Elo system's speed in all cases.

The buffering process seems to overestimate the the rating of better player and underestimate that of the weaker player. This is more apparent when the system is tested for stability by starting the players at their true ratings and observing the changes.



The system appears to have good stability for the player pair, when compared to the Elo system. The results from the stability tests mentioned in the methodology follow *i.e.* for each player pair, start both players with assigned ratings equal to true ratings, simulate $10\,000$ games and calculate the average assigned rating and standard rating deviation. An assigned rating mean close to the true rating mean accompanied by a small standard rating deviation indicates a stable system.

True Rating	Difference	Elo	Elo SD	Buffer	Buffer SD
k = 10					
Player A					
1550	100	1550	22	1558	24
1600	200	1599	20	1614	23
1700	400	1700	20	1721	20
1800	600	1800	18	1823	20
1900	800	1898	15	1910	23
2000	1000	1999	11	2001	19
Player B					
1000	1000	1001	11	999	19
1100	800	1102	15	1090	23
1200	600	1200	18	1177	20
1300	400	1300	20	1279	20
1400	200	1401	20	1386	23
1450	100	1450	22	1442	24
k = 15					
Player A					
1550	100	1551	25	1560	31
1600	200	1603	24	1613	30
1700	400	1702	25	1719	29
1800	600	1810	23	1821	25
1900	800	1913	20	1909	20
2000	1000	1999	18	2030	15
Player B					
1000	1000	1001	18	970	15
1100	800	1087	20	1091	20
1200	600	1190	23	1179	25
1300	400	1298	25	1281	29
1400	200	1397	24	1387	30
1450	100	1449	25	1440	31
k = 16					
Player A					
1550	100	1553	27	1558	34
1600	200	1599	28	1613	31
1700	400	1704	28	1714	29
1800	600	1796	25	1828	27

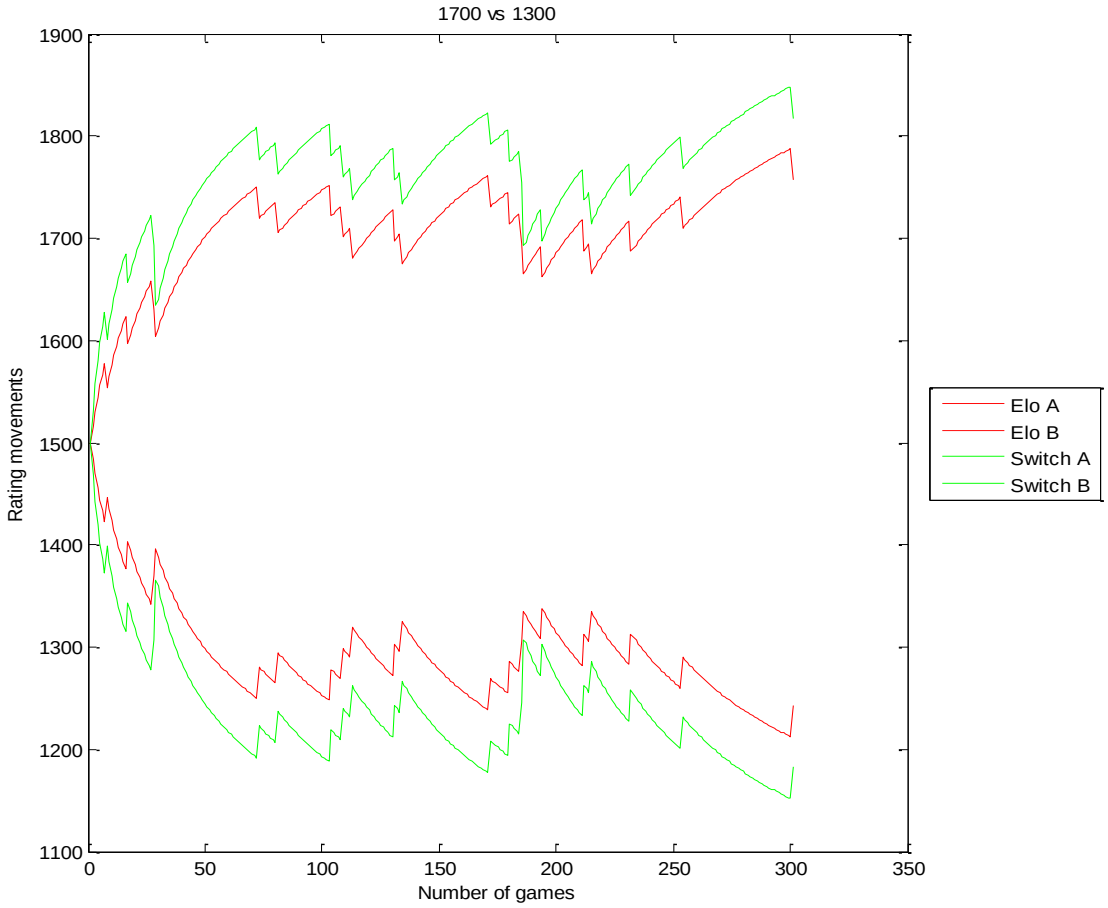
1900	800	1888	25	1904	23
2000	1000	2008	18	2033	13
Player B					
1000	1000	992	18	967	13
1100	800	1112	25	1096	23
1200	600	1204	25	1172	27
1300	400	1296	28	1286	29
1400	200	1401	28	1387	31
1450	100	1447	27	1442	34
k = 24					
Player A					
1550	100	1554	34	1558	36
1600	200	1603	33	1618	36
1700	400	1705	31	1718	32
1800	600	1806	29	1812	34
1900	800	1900	28	1908	30
2000	1000	1996	29	2002	34
Player B					
1000	1000	1004	29	998	34
1100	800	1100	28	1092	30
1200	600	1194	29	1188	34
1300	400	1295	31	1282	32
1400	200	1397	33	1382	36
1450	100	1446	34	1442	36
k = 25					
Player A					
1550	100	1550	34	1561	39
1600	200	1601	34	1617	38
1700	400	1706	30	1724	31
1800	600	1799	33	1824	31
1900	800	1910	29	1928	30
2000	1000	1998	34	2043	24
Player B					
1000	1000	1002	34	957	24
1100	800	1090	29	1072	30
1200	600	1201	33	1176	31
1300	400	1294	30	1276	31
1400	200	1399	34	1383	38
1450	100	1450	34	1439	39
k = 32					
Player A					
1550	100	1554	37	1561	44
1600	200	1602	36	1617	41
1700	400	1703	34	1725	39
1800	600	1807	40	1832	37
1900	800	1900	28	1927	28

2000	1000	2011	41	2030	34
Player B					
1000	1000	989	41	970	34
1100	800	1100	28	1073	28
1200	600	1193	40	1168	37
1300	400	1297	34	1275	39
1400	200	1398	36	1383	41
1450	100	1446	37	1439	44

In the above table, the column labelled 'difference' is the difference between the two players' ratings used in the specific simulation. The columns labelled 'Elo' and 'Buffer' give the average assigned ratings by the Elo and Buffer systems respectively, while the columns labelled 'Elo SD' and 'Buffer SD' give the respective standard deviations of the assigned ratings. From the table it can be seen that the Buffer system always overestimates the rating of the stronger player and underestimates the rating of the weaker player. This also occurs in the case of the Elo system, but the estimation error of the Buffer system is larger in all cases. The standard deviations of the assigned ratings are also higher in the case of the Buffer system, except for large k-values accompanied by large differences in true ratings.

The Switching Momentum system

A different approach to the momentum addition would be to alter the k-value used in the rating adjustment equation under different streak and non-streak situations. Intuitively, one would argue to use a higher k-value for a player who is on a streak, but not indefinitely. As a proof of concept, a system is suggested where a player has a momentum of -1, 1 or 0. A momentum of 0 indicates that the player is currently not on a streak, whereas a momentum of 1 or -1 indicates a winning and losing streak respectively. A streak is defined as winning (or losing) 2 games in a row. As with the buffer system, we start by testing the system for two players with ratings 1300 and 1700 and plot the rating changes.



It is immediately apparent that the assigned ratings under the Switching Momentum system reach the true ratings faster than those of the Elo system. To verify this, we perform the test for the system's speed as described in the methodology; that is, we simulate games under both systems to calculate the average number of games taken to reach the true rating difference. We then do this for all the above mentioned player and k-value combinations. Each time, we do a t-test of equal means to test the following hypothesis:

$$H_0: \mu_{Elo} = \mu_{switching\ momentum}$$

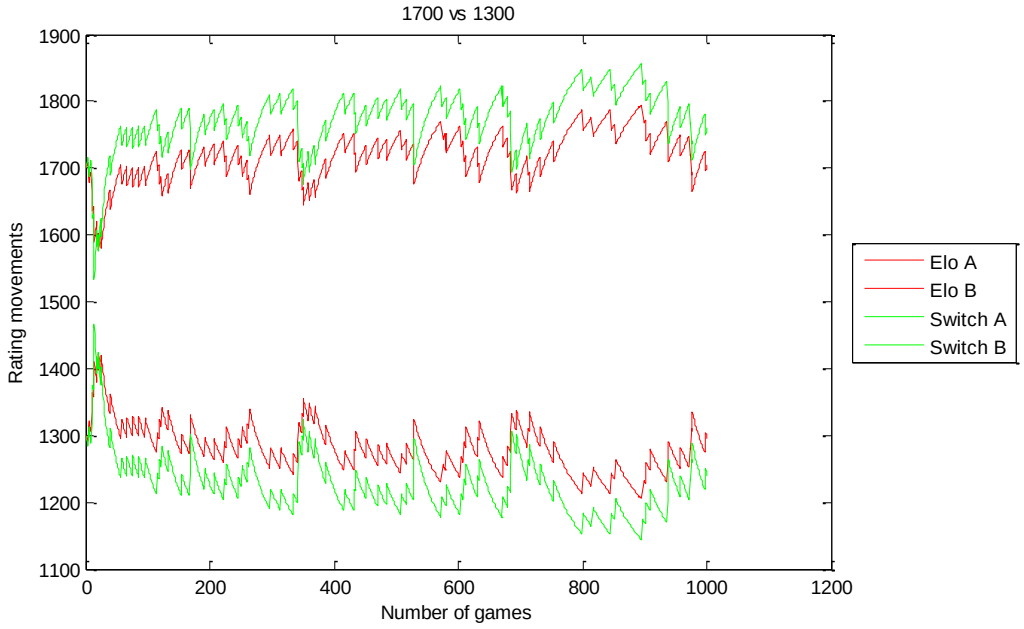
where μ_i is the average number of games taken for the assigned ratings to reach the true ratings under each system.

Rating diff	**k= 10**		**k= 15**		**k= 16**		**k= 24**		**k= 25**		**k= 32**	
	Elo	**Switch**	**Elo**	**Switch**	**Elo**	**Switch**	**Elo**	**Switch**	**Elo**	**Switch**	**Elo**	**Switch**
100												
Mean	62	21	37	12	34	11	20	6	19	6	13	4
SD	40	15	26	9	24	8	15	4	14	4	10	3
200												
Mean	100	32	61	21	57	19	35	12	33	12	24	9
SD	48	14	32	10	29	9	19	7	18	7	14	5

400												
Mean	243	75	149	49	138	45	85	29	81	28	60	22
SD	98	18	62	13	57	13	37	9	36	9	27	7
600												
Mean	668	204	414	133	382	125	237	81	224	77	166	60
SD	258	40	171	29	156	27	100	20	94	19	71	16
800												
Mean	1999	606	1230	398	1131	371	696	242	662	233	494	179
SD	785	111	493	83	454	80	286	58	271	55	209	46
1000												
Mean	6186	1864	3814	1230	3563	1152	2169	749	2065	716	1524	553
SD	2399	327	1523	251	1470	244	898	174	861	166	633	138

To determine whether to use the assumption of equal variances, F-tests were completed in each case, and every time the hypothesis of equal variances was rejected at a confidence level of 95%. Thus, a t-test assuming unequal variances was used for each pair. In each case the test showed a significant difference between the average numbers of games taken to reach the true rating difference. From this it can be concluded that the Switching Momentum system has more speed than the Elo system by the before-mentioned standard. Overall, the Switching Momentum system is almost 3 times faster than the basic Elo. It is also interesting to note that the Switching system with a k-value of 10 had a performance that can be compared to the Elo system with a k-value of 24.

The analysis continues with a look at the stability of the Switching Momentum system when compared with the Elo system. Both players start with their assigned ratings equal to their actual ratings, 1300 and 1700, the changes are logged and plotted.



Though the speed of the Switching Momentum system is better than the Elo system, it is clear that it is not as stable. This is intuitive, since the k-value used is doubled every time a player wins two games in a row. The results of the stability tests from all the player and k-value combinations follow. The table shows the true ratings of the players for each k-value, the difference between the ratings, then the average rating assigned by the both systems, as well as a standard deviation for this rating.

True Rating	Difference	Elo	Elo SD	Switch	Switch SD
k = 10					
Player A					
1550	100	1549	21	1569	36
1600	200	1600	20	1635	31
1700	400	1701	21	1751	26
1800	600	1795	15	1852	17
1900	800	1901	14	1956	20
2000	1000	1993	11	2039	17
Player B					
1000	1000	1007	11	961	17
1100	800	1099	14	1044	20
1200	600	1205	15	1148	17
1300	400	1299	21	1249	26
1400	200	1400	20	1365	31
1450	100	1451	21	1431	36
k = 15					
Player A					
1550	100	1552	25	1574	42
1600	200	1603	25	1641	38
1700	400	1705	24	1759	28
1800	600	1811	25	1869	28
1900	800	1915	23	1972	25
2000	1000	2001	19	2051	20
Player B					
1000	1000	999	19	949	20
1100	800	1085	23	1028	25
1200	600	1189	25	1131	28
1300	400	1295	24	1241	28
1400	200	1397	25	1359	38
1450	100	1448	25	1426	42
k = 16					
Player A					
1550	100	1553	27	1575	47
1600	200	1599	27	1636	42
1700	400	1701	28	1753	33
1800	600	1792	22	1849	24
1900	800	1886	26	1942	29

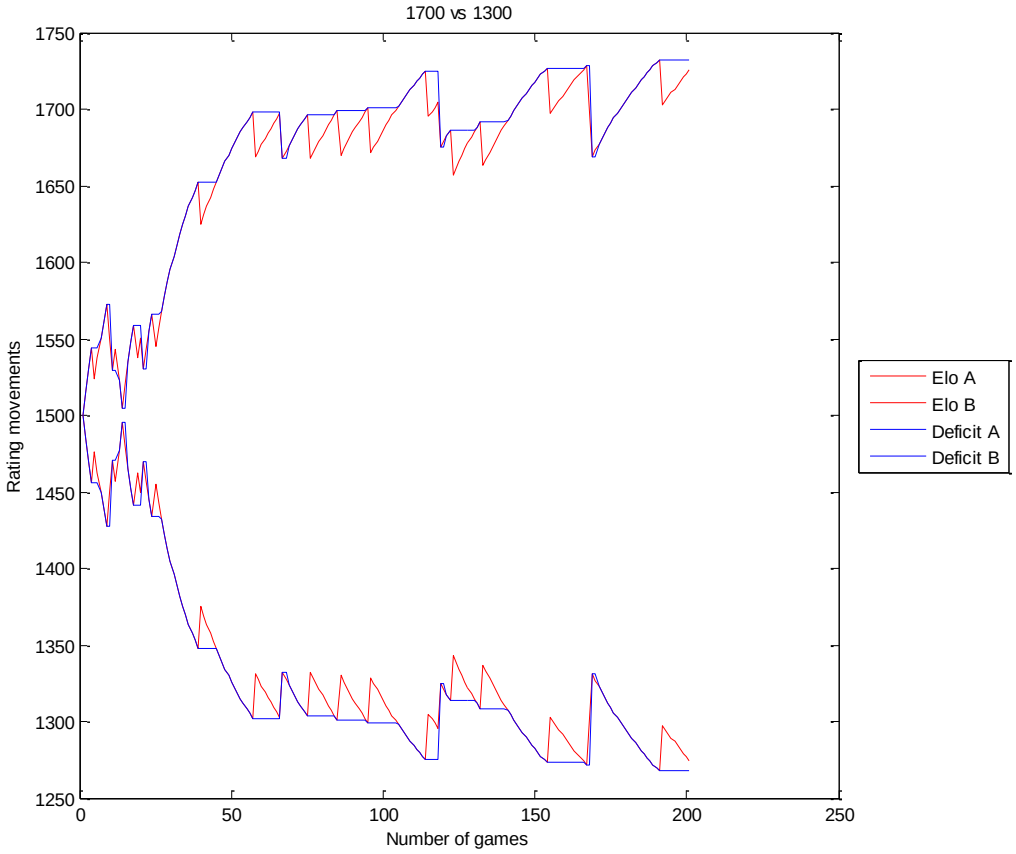
2000	1000	2006	19	2057	23
Player B					
1000	1000	994	19	943	23
1100	800	1114	26	1058	29
1200	600	1208	22	1151	24
1300	400	1299	28	1247	33
1400	200	1401	27	1364	42
1450	100	1447	27	1425	47
k = 24					
Player A					
1550	100	1553	33	1577	57
1600	200	1603	33	1642	50
1700	400	1706	31	1759	37
1800	600	1810	30	1868	32
1900	800	1905	27	1963	28
2000	1000	2002	27	2057	29
Player B					
1000	1000	998	27	943	29
1100	800	1095	27	1037	28
1200	600	1190	30	1132	32
1300	400	1294	31	1241	37
1400	200	1397	33	1358	50
1450	100	1447	33	1423	57
k = 25					
Player A					
1550	100	1550	34	1572	58
1600	200	1601	34	1639	51
1700	400	1706	31	1760	36
1800	600	1801	33	1858	35
1900	800	1907	27	1965	29
2000	1000	2007	35	2060	41
Player B					
1000	1000	993	35	940	41
1100	800	1093	27	1035	29
1200	600	1199	33	1142	35
1300	400	1294	31	1240	36
1400	200	1399	34	1361	51
1450	100	1450	34	1428	58
k = 32					
Player A					
1550	100	1553	38	1576	64
1600	200	1603	36	1643	53
1700	400	1704	34	1759	39
1800	600	1805	39	1862	41
1900	800	1902	33	1961	35
2000	1000	2009	42	2066	45

Player B					
1000	1000	991	42	934	45
1100	800	1098	33	1039	35
1200	600	1195	39	1138	41
1300	400	1296	34	1241	39
1400	200	1397	36	1357	53
1450	100	1447	38	1424	64

A stable system will have an average assigned rating that is close to the real rating, as well as a small standard deviation. From the graph and table it is immediately obvious that the Switching Momentum system does not have the accuracy or stability of the Elo system. This is true for all k-values. For larger true rating differences the Standard deviation of the Switching Momentum system isn't much larger than that of the Elo, but it is offset by the average assigned rating being inaccurate. This system tends to greatly overestimate the better player and underestimate the weaker player. At this stage it is clear that using momentum to alter the k-value used in calculations can greatly increase the speed of the Elo system, even for a simple case with momentum only having three possible values. It is only the stability that should be improved on.

The Deficit system

In our final momentum-added adjustment of the Elo system, we track the ratings assigned by the Elo system, but create a "deficit" when a streak is broken. A player's rating will remain unchanged while playing out of a negative or positive deficit, possibly ensuring greater stability. Higher rated players will not experience a large decline in their assigned rating following an accidental loss against a much lower rated player. Instead, a deficit will be created and the player's assigned rating should remain unchanged if the player can play themselves out of the deficit, which should be the case if the loss was accidental rather than due to a true difference in skill. The Deficit system can also be adapted to assign higher k-values on winning or losing streaks, but then it will not track the Elo system. Doing the test for speed and plotting the results for the 1300 – 1700 player pair, yields the following.



Because the Deficit system tracks the Elo system, the speed is exactly the same. We test the speed by using the hypothesis described in the methodology, for all the different player and k-value combinations:

$$H_0: \mu_{Elo} = \mu_{Deficit}$$

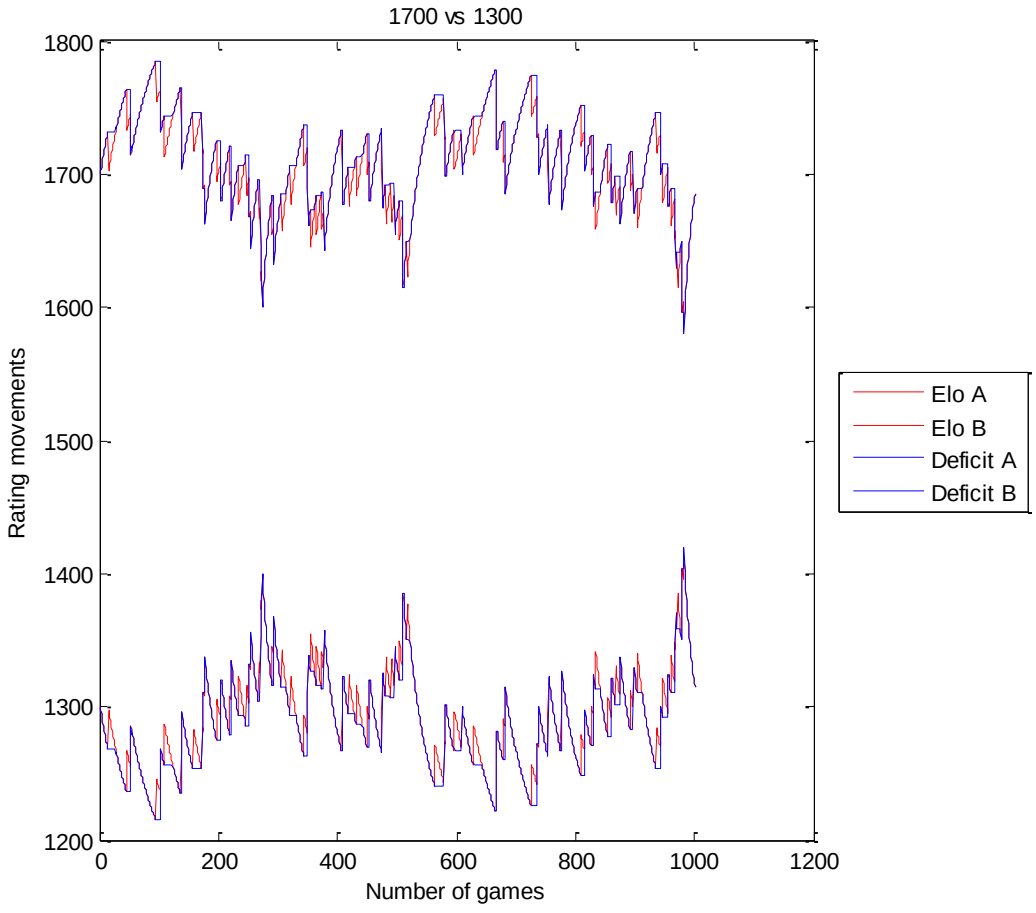
where μ_i is the average number of games taken for the assigned ratings to reach the true ratings under each system.

Rating difference	k = 10		k = 15		k = 16		k = 24		k = 25		k = 32	
	Elo	Deficit	Elo	Deficit	Elo	Deficit	Elo	Deficit	Elo	Deficit	Elo	Deficit
100												
Mean	62	62	37	37	34	34	19	19	18	18	13	13
SD	40	40	26	26	24	24	15	15	14	14	10	10
200												
Mean	100	100	61	61	57	57	35	35	33	33	24	24
SD	48	48	31	31	29	29	19	19	18	18	14	14
400												
Mean	243	243	148	148	138	138	85	85	81	81	60	60
SD	98	98	62	62	57	57	37	37	36	36	27	27
600												
Mean	668	668	415	415	382	382	237	237	224	224	166	166
SD	258	258	171	171	157	157	99	99	94	94	71	71

800												
Mean	2000	2000	1229	1229	1131	1131	696	696	662	662	494	494
SD	786	786	492	492	453	453	286	286	271	271	208	208
1000												
Mean	6192	6192	3800	3800	3543	3543	2166	2166	2063	2063	1529	1529
SD	2448	2448	1527	1527	1428	1428	895	895	858	858	634	634

F-tests were performed for all of the ratings to determine whether the assumption of equal variances holds. At the 95% confidence level, the hypothesis of equal variance was rejected for all of the above, so t-tests using unequal variances were performed. In each case the test showed an insignificant difference between the average numbers of games taken to reach the true rating difference. Thus, it can be concluded that the speed of the Deficit system is statistically the same as that of the Elo system. This was expected, since the Deficit system tracks the Elo system when the k-factor is not inflated on streaks, as was the case of the above tests.

Stability is again checked by starting both players at their true ratings and producing a plot of the changes in the assigned ratings.



From looking at the graph, first impressions would suggest that the Deficit system has stability close to the Elo system. The following table shows the results of the stability test, as described in the methodology. Again the table shows the true ratings of the players for each k-value, the difference between the ratings, then the average rating assigned by the both the Elo and Deficit systems, as well as a standard deviation for this rating.

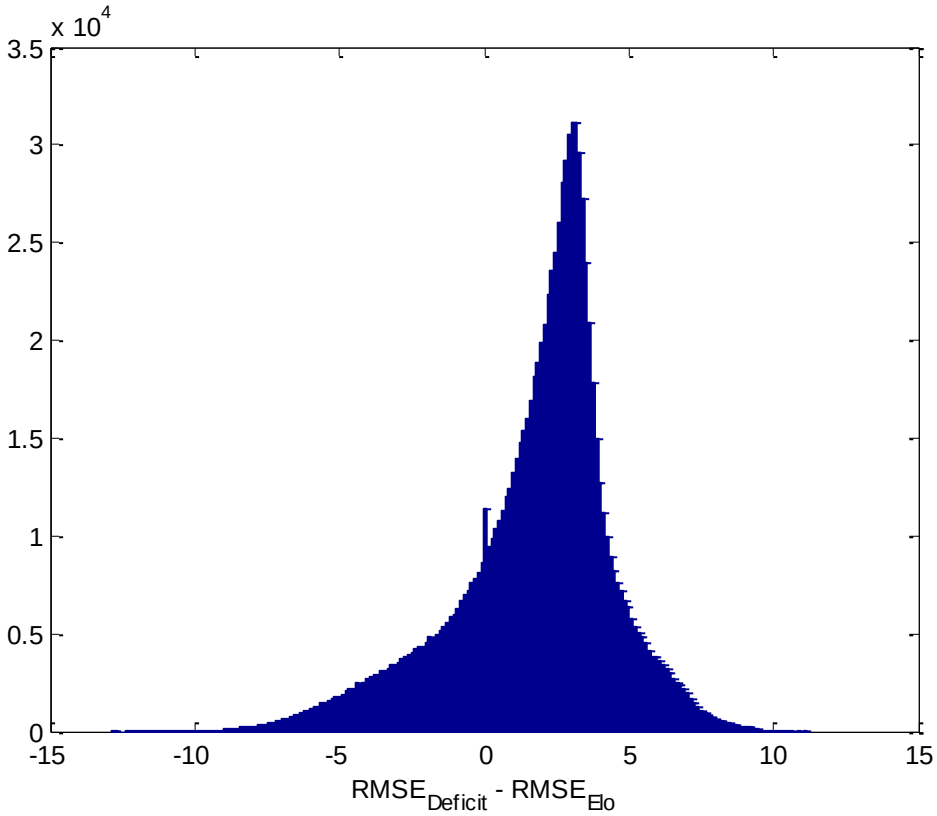
True Rating	Difference	Elo	Elo SD	Buffer	Buffer SD
k = 10					
Player A					
1550	100	1550	22	1550	22
1600	200	1599	20	1599	20
1700	400	1700	20	1700	21
1800	600	1800	18	1800	18
1900	800	1898	15	1899	15
2000	1000	1999	11	2000	11
Player B					
1000	1000	1001	11	1000	11
1100	800	1102	15	1101	15
1200	600	1200	18	1200	18
1300	400	1300	20	1300	21
1400	200	1401	20	1401	20
1450	100	1450	22	1450	22
k = 15					
Player A					
1550	100	1551	25	1551	26
1600	200	1603	24	1603	25
1700	400	1702	25	1702	25
1800	600	1810	23	1810	24
1900	800	1913	20	1914	20
2000	1000	1999	18	2000	18
Player B					
1000	1000	1001	18	1000	18
1100	800	1087	20	1086	20
1200	600	1190	23	1190	24
1300	400	1298	25	1298	25
1400	200	1397	24	1397	25
1450	100	1449	25	1449	26
k = 16					
Player A					
1550	100	1553	27	1553	27

1600	200	1599	28	1599	28
1700	400	1704	28	1704	29
1800	600	1796	25	1796	26
1900	800	1888	25	1888	25
2000	1000	2008	18	2009	19
Player B					
1000	1000	992	18	991	19
1100	800	1112	25	1112	25
1200	600	1204	25	1204	26
1300	400	1296	28	1296	29
1400	200	1401	28	1401	28
1450	100	1447	27	1447	27
k = 24					
Player A					
1550	100	1554	34	1554	36
1600	200	1603	33	1603	34
1700	400	1705	31	1705	32
1800	600	1806	29	1806	29
1900	800	1900	28	1900	29
2000	1000	1996	29	1996	29
Player B					
1000	1000	1004	29	1004	29
1100	800	1100	28	1100	29
1200	600	1194	29	1194	29
1300	400	1295	31	1295	32
1400	200	1397	33	1397	34
1450	100	1446	34	1446	36
k = 25					
Player A					
1550	100	1550	34	1550	36
1600	200	1601	34	1602	36
1700	400	1706	30	1706	31
1800	600	1799	33	1800	34
1900	800	1910	29	1910	29
2000	1000	1998	34	1999	35
Player B					
1000	1000	1002	34	1001	35
1100	800	1090	29	1090	29
1200	600	1201	33	1200	34
1300	400	1294	30	1294	31
1400	200	1399	34	1398	36
1450	100	1450	34	1450	36

k = 32					
Player A					
1550	100	1554	37	1554	40
1600	200	1602	36	1603	38
1700	400	1703	34	1703	35
1800	600	1807	40	1808	41
1900	800	1900	28	1900	28
2000	1000	2011	41	2012	43
Player B					
1000	1000	989	41	988	43
1100	800	1100	28	1100	28
1200	600	1193	40	1192	41
1300	400	1297	34	1297	35
1400	200	1398	36	1397	38
1450	100	1446	37	1446	40

The average rating assigned by the Deficit system seems just as good as that of the Elo system. The deviation of the Deficit system's assigned rating is very close to that of the Elo system, especially for small k-factor values. For the higher k-factor values, the Deficit system has slightly higher deviations than the Elo system and it is worth investigating further by means of the RMSE method mentioned in the methodology.

Analysing the difference between the RMSE values yields the following histogram and characteristics:



- Mode 0
- Median 2.33
- Mean 1.77

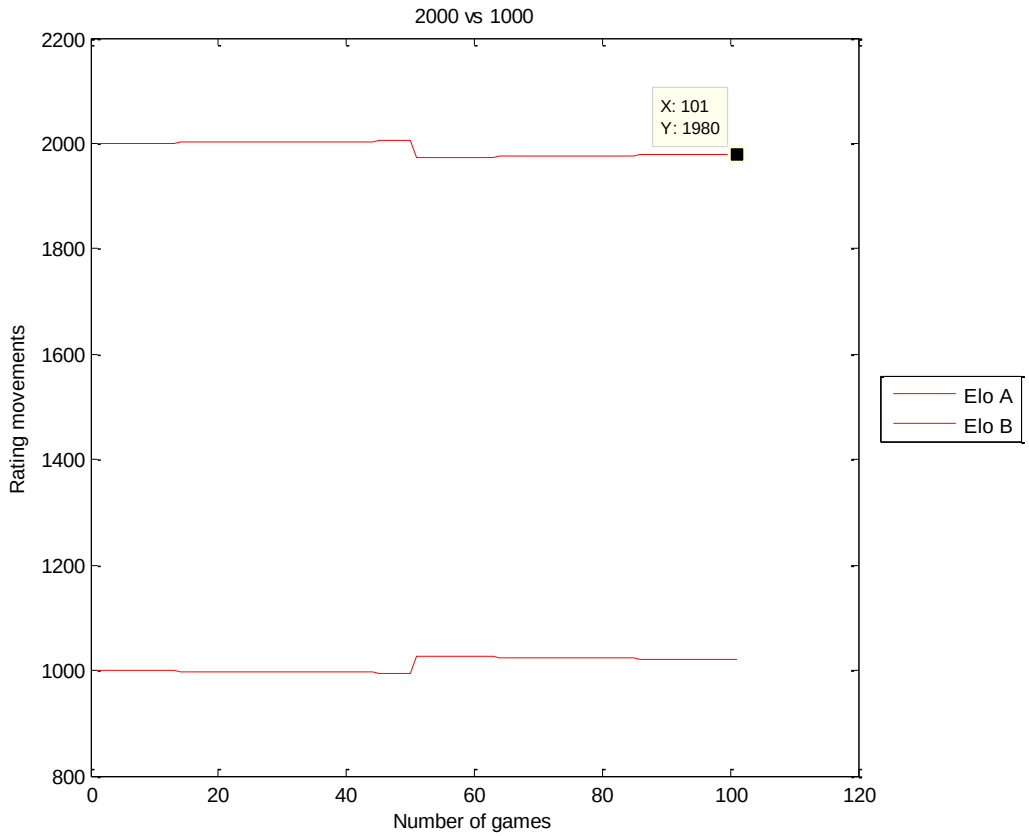
Percentiles:

2.50%	5%	50%	95%	97.50%
-5.02	-3.76	2.34	5.64	6.48

It is clear that the differences between the RMSE's are of a small magnitude. The percentiles span zero, showing that many of the differences were insignificant. The median and mean are also of a very small magnitude; In fact, the mode is equal to zero, because of the rounding properties of the software used – when the true difference between the players is very high, the probability of the higher-rated player losing is very small, resulting in the systems having exactly the same rating adjustments and thus the same RMSE's. From these results it can be concluded that the stability of the Deficit system is very close to the Elo.

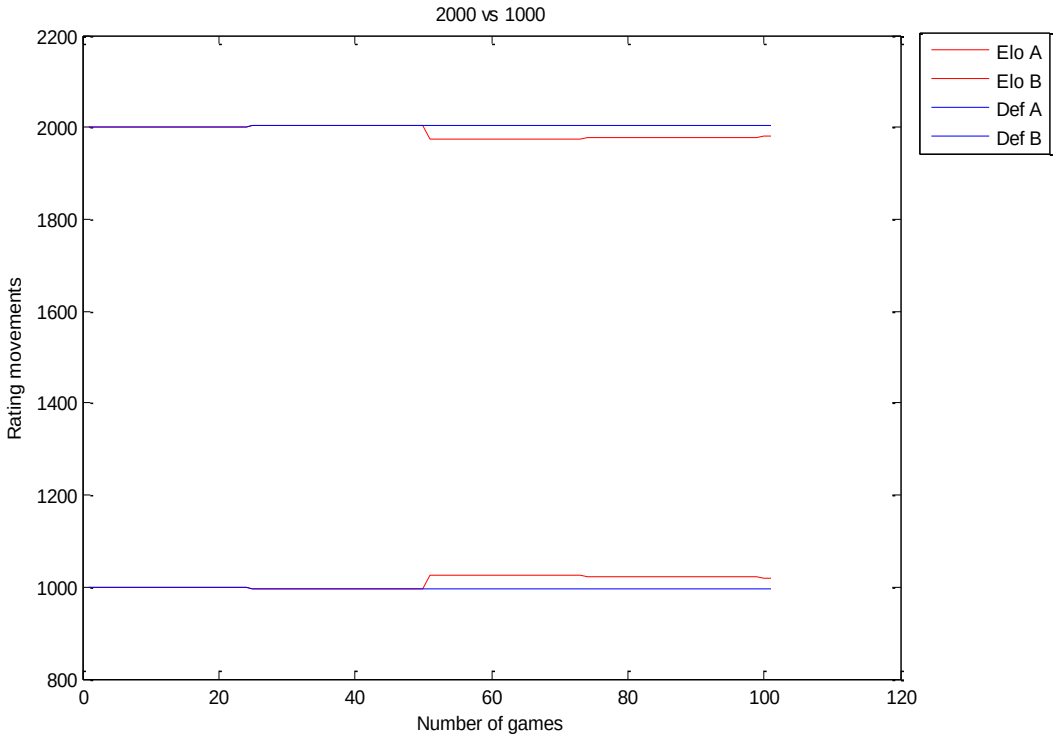
It is worth mentioning a shortcoming of the basic Elo system, which is the effect of an unnatural loss of a very highly rated player against player with a much lower rating. An unnatural loss can be regarded as the higher-rated player losing due to reasons other than skill *i.e.* an accidental loss. This might be an unlikely scenario in chess tournaments, but it could easily occur in the arena of internet

gaming. In this case the assigned ratings for both players would undergo an immense change, which could take a long time to be corrected. This is illustrated by simulating 100 games between two players, with ratings 1000 and 2000, and forcing the higher player to lose halfway through. The following plot is produced by rating the players using the Elo system.



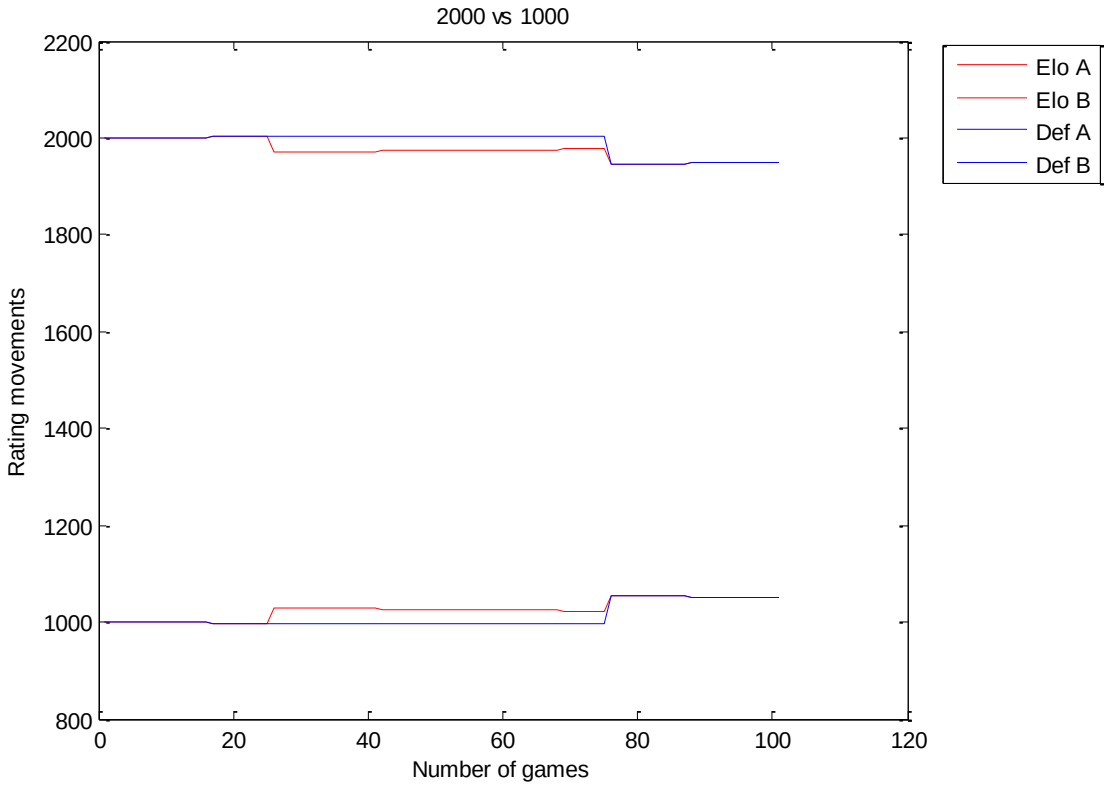
RMSE = 17

After the unnatural loss, there is a large change in the ratings, which is corrected very slowly. Fifty games after the unnatural loss, the ratings still haven't recovered. The Deficit system addresses this problem in the use of the momentum component. The deficit created for both players puts them on 'probation,' rather than immediately changing the ratings. The following plot shows the same scenario as above, comparing the rating movements of the Elo and Deficit systems.



RMSE:
Elo 17
Deficit: 4

Momentum acquired by both players in the first 50 games protected the players from the unnatural event at the 50th game. After this, all of the ratings assigned by the Deficit system can be regarded as more accurate than that of the Elo system. Next, a scenario is considered where the higher player loses after 25 games and again at 75 games.



RMSE:
Elo 33
Deficit: 27

Even though the rating changes at the 25th game were prevented by the momentum held by both players, the second loss at the 75th game suggests that the earlier loss might not have been entirely unnatural. Thus, the rating is adjusted to where it would have been, ignoring momentum. This illustrates the probation period of the Deficit system; while a player is playing out of a Deficit, he/she is not protected by momentum.

Conclusion

This paper explored different ways of incorporating momentum with the aim to make the Elo system more efficient. The Buffer system used sequential momentum build up to protect against later rating changes. Another approach was the Switching momentum system, which altered the k-factor used when on streaks. Both of these systems improved on the speed at which true ratings are obtained, at the expense of stability of the assigned ratings.

The Deficit system tracks the Elo system and creates deficits when a streak is broken, halting all rating movements until the deficit is recovered, or the initial unnatural break is proved to be a

natural one. This system had stability almost in line with the Elo system, while addressing one of its flaws. From this it could be stated that the Deficit system is more effective than the Elo system, as it has the same strengths minus one of the weaknesses. The addition of momentum, in all three of the above mentioned systems, didn't complicate calculations to a large extent. All three of the tested systems required only the current rating and the current momentum to be recorded for players, as opposed to only recording the current rating as is the case with the Elo system. Rating adjustments can be done without the need of considerable computing powers and players will be able to calculate rating changes themselves. This is important to ensure transparency of the systems.

In conclusion, it was found that momentum could be used to make the basic Elo rating system more effective.

Problems with adding momentum

Adding a momentum component to ratings, can lead to adverse selection. Large negative holdings of momentum could affect the morale of players, to such an extent that they might drop out of tournaments. The same could apply to players with high momentum holdings, adverse selection could occur where players choose only to enter some tournaments that seems beneficiary to them. Negative momentum holdings could also be an unattractive prospect to new rating environments, making it hard to implement the system in practice.

Continued adapting of systems using momentum could result in a loss of simplicity, which could also hinder the process of implementing rating systems to new environments outside of chess.

Further research

This paper only provided a brief outline of possible methods to implement momentum to create rating systems based on the Elo system, endeavouring to make it more effective. There remains space to thoroughly test these systems with various changes the parameters used, for example, the Deficit system could be changed to adjust the k-factor on streaks. In this paper, only two player environments were considered. It is of interest to investigate how these systems would respond in tournament environments, or independent groups of players with one or two players interchanging. The latter would give insight to how the ratings of the two independent groups vary relative to the ratings in the other group.

It is important to note that the system Elo system, as well as the adjusted systems proposed, hold great potential for measuring team games and games with more than two players (or teams), and is additionally not restricted to win, draw or loss outcomes, but can be used in rating systems with score-based games as well. An analysis of the Elo and the proposed ratings systems under these conditions should be carried out in future.

The rating systems could also be used to rate games of archived tournaments to see how the outcome would have been affected. Real world application to recreational environments could also be considered to identify possible problems caused by social dynamics.

References

Coulom, R., 2008. Whole-History Rating: A Bayesian Rating System for Players of Time-Varying Strength. *Computers and Games,* 6,pp. 113-124.

Elo, A.E., 1978. *The rating of chessplayers: past and present*. (2008 ed). New York: Ishi Press International.

Glickman, M.E., 1995. A comprehensive Guide to Chess Ratings.*American Chess Journal,* 3, pp.59-102.

Glickman, M.E. & Jones, A.C., 1999.Rating the chess rating system.*Chance*, 12(2), pp. 21-28.

MATLAB version 7.5.0.342, Natick, Massachusetts: The MathWorks Inc., 2007.

Ross, D., 2007. Arpad Elo and the Elo rating system.*Chessbase News,* [internet] 16 December.
Available at: http://www.chessbase.com/newsdetail.asp?newsid=4326
[Accessed 28 February 2010].

Sloan, S., 2008.Introduction. In: Elo, A.E., ed. 2008. *The rating of chessplayers: past and present.* New York: Ishi Press International, pp. 5-16.

Vovk, B, 2008, Chess Elo Rating in Simple. *www.chesselo.com.* [internet] (Updated 9 December 2010)
Available at: http://www.chesselo.com/index.html
[Accessed 11 December 2010]