ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Silesian Statistical Review

2024, nr 22(28) ISSN 2449-9765

DOI: 10.15611/sps.2024.22.05

Evaluation of the Elo Rating System Based on the Results of the 44th Olympics in Chennai

Katarzyna Ostasiewicz

Wroclaw University of Economics and Business

e-mail: katarzyna.ostasiewicz@ue.wroc.pl

ORCID: 0000-0002-0115-3696

Quote as: Ostasiewicz, K. (2024). Evaluation of the Elo Rating System Based on the Results of the 44th Olympics in Chennai. *Silesian Statistical Review*, *22*(28), 39-49.

JEL: C10, C52, C63

Abstract: The characteristics of a professional chess player is described by the so-called Elo rating which is supposed to reflect the strength of the player, and is updated based on his or her performance, i.e. their wins, losses and draws with other players. A win or draw with higher-rated player means the increase in the rating of the given player, while a loss or draw with lower-rated player decreases his or her rating position. These gains and losses in the rating system are calculated according to a specific algorithm. The Elo rating is supposed to predict the expected score of a player having a given advantage or disadvantage over his or her opponent. The question is, however, how good these predictions are? Therefore, we want to investigate the accuracy of the Elo model, based on the results of the 44th Olympics in Chennai.

Keywords: Elo model, chess rating

1. The Elo Model

According to the Elo model (Elo, 1978), the strength of play of each player is described by normal distribution with some expected value, which is his or her Elo rating, R. Thus, the performance of each player is not fully determined by his or her strength but can fluctuate in a random way, and it is possible for the weaker player to win with the player with higher rating. Still, the greater the difference between the strengths of the players is, the less possible is that weaker player will outperform the stronger one. Specifically, if player A with rating R_A plays with player B with rating R_B, the expected score for player A in many games is

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

See also Fig. 1 with difference of ratings (diff) on the horizontal axis and expected score on the vertical one.

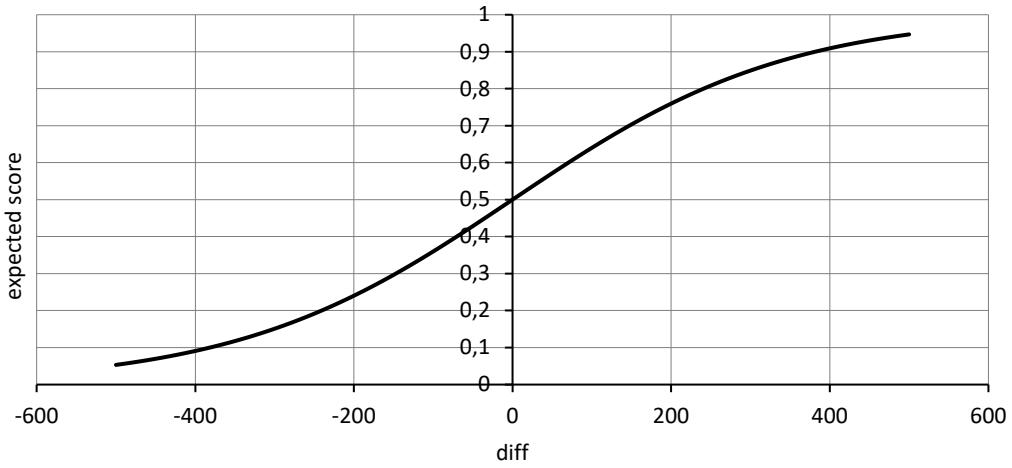


Figure 1. Expected score of a player having an advantage over the opponent of a number of Elo points denoted as diff. Negative value means disadvantage of the player

Source: own elaboration.

For equal ratings $R_A = R_B$, the expected score is equal to ½, and the greater the advantage of A over B is, the higher the expected score of A is. For a difference equal to 400, the stronger player is expected to score almost 91% of available points. Note that for $R_A < R_B$ the advantage is negative, which is the same as disadvantage. Note also that within the Elo model, there is no difference between two draws (average result equals ½) and one win and one loss (average result equals again ½).

The 44th Olympics in Chennai collected players from all over the world with very different ratings, and all games took place only in a few days. Thus, it is supposed that the strength of a given player did not change during this event (although their performance could change according to normal distribution underlying the expected value of their strength). 'An individual' here is not a particular player, but a group of all players of the same rating. We can expect that within such a group, some players will have 'a good day' and some – worse; thus, the performance of players of the same rating will be distributed randomly with the average value equal to their common rating.

Although there are many other proposals to measure the strength of chess players (Glickman & Jones, 1999; Sismanis, 2010; Veček et al., 2014), Elo model remains the best-known and easiest to apply in practice.

Apart from checking the accuracy of the Elo model predictions, we will also estimate the influence of playing white. According to the Elo formula, the expected score of a player depends only on his or her advantage or disadvantage in the rating over their opponent. However, it is strongly believed that playing white gives additional advantage, e.g. Stockfish chess engine from the very beginning gives the white player such an advantage as if she/he had a half of pawns extra. We will check this potential advantage basing on the obtained data.

2. The Elo constant

The first step is to compare the empirical data with the Elo model. Figure 2 presents the average score of players whose advantage over their rivals was equal to a given value. Those values were grouped into intervals of the width of 20 rating points to obtain the frequencies for which it would

be reasonable to use the laws of the calculus of probability. Empirical average scores (solid line) were compared with the expectations of the Elo model (dashed line). It can be easily observed that nearly the entire empirical curve lies beneath the dashed line, i.e. the Elo model overestimates the real results of players.

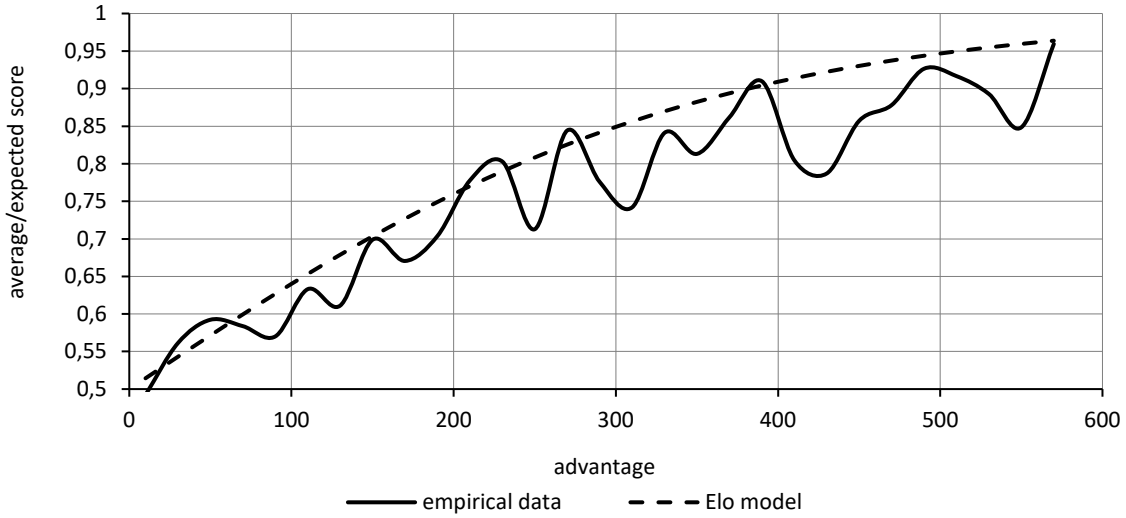


Figure 2. Average score of players with the given advantage over the rival as compared with the predictions of the Elo model with the constant 400

Source: own elaboration.

It may be supposed that by manipulating the constant in the Elo model, one may obtain a better fit to empirical data.

In what follows, we will use logistic model rather than raw empirical data. Why? The very nature of empirical data is that they are finite in number and influenced by randomness (clearly visible in Fig. 2). Statistical methodology, however, allows us to obtain a smooth curve that is best fitted to the empirical data. Additionally, the statistical method informs us whether the picture that emerges after eliminating randomness may be treated as a real picture of the investigated phenomena (basing on *p*-values). Having such a model, we can make predictions also for the values which have not been observed within empirical data.

Thus, we will apply the logistic model for the expected value of wins and draws, taking the difference between the given player and his or her rival as an explanatory variable. Fortunately, the logistic function has the same formula as the function used by Elo, with the only difference being the fact that the basis of the power within the logistic function is e instead of 10:

$$p(\bar{x}) = \frac{1}{1+e^{-\bar{\alpha}\cdot\bar{x}}}, \tag{2}$$

where $\bar{x}$ denotes the whole possible set of explanatory variables with coefficients $\bar{\alpha}$. Still, these formulas can be easily transformed one into another by:

$$e^{Y} = 10^{\log_{10} e^{Y}} = 10^{Y\log_{10} e}, \tag{3}$$

and

$$10^{Y} = e^{\ln 10^{Y}} = e^{Y\ln 10}. \tag{3a}$$

Estimation of the logistic model for the expected score with the difference between players as the explanatory variable gives values shown in Tab. 1.

Table 1. The results of the model for the expected score of a player having an advantage equal to the 'difference' over his opponent

	Estimate	Standard error	*z*-statistics	*p*-value
Constant	–4.558*10^(–16)	0.01952	–2.335*10^–14	1.
Difference	0.00449	0.0000875355	51.3458	5.078*10^(–575)

Source: own elaboration.

The constant is insignificant; thus, it can be put as zero. The difference is highly significant with parameter 0.00449. This value inserted into (3) gives 0.00195. As in the Elo model the constant is the inverse of 400, let us also calculate the inverse of the calculated value, which is about 512 (precisely, 512.303), cf. (Glickman & Jones, 1999). It seems that the actual data would be better described by the modified Elo model with the constant 1/400 substituted by the constant 1/512 – see the dotted line added in Fig. 3.

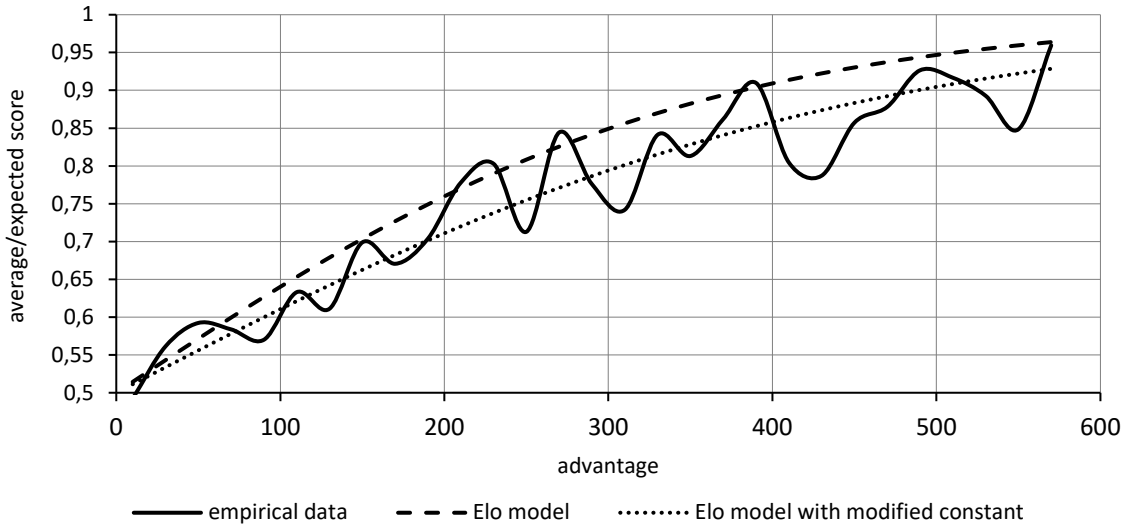


Figure 3. Average score of players with a given advantage over their rival as compared with the Elo model predictions with the constant 400 and with the Elo model with the modified constant equal to 512

Source: own elaboration.

Maybe the Elo model with the constant 400 is better fitted for couples of players with similar ratings? If we restrict the data to the couples with differences not exceeding 400, the constant of the model appears to be equal to 486. If we further restrict it to differences not exceeding 200, we will get 518. Thus, it is not the right direction.

Now then, maybe the Elo model with the constant 400 is better fitted for players with high ratings? To check this possibility, we have estimated logistic models for data restricted to couples of players whose both ratings are within the range (1500 + Δ, 1900 + Δ). That means that we started from couples of two relatively weak players, whose ratings were in the range (1500, 1900). Then, we shifted to couples of players who were a bit better, both from the (1510, 1910) interval, and so on, keeping all the time the interval width of 400 points, but moving towards better and better players. We started from the interval (1500, 1900) as the couples of lower-rated players were too few to estimate the logistic model. Having estimated the models, we could calculate the corresponding 'modified Elo constant,' i.e. the quantity that stands in the resulting fitted model for the original value 400 in the Elo model. The resulting constants turned out to differ a lot, which is visualised in Fig. 4.

Figure 4. Modified Elo constant vs the middle of the range of the rating of players

Source: own elaboration.

Added dashed line smooths the data a bit (running average with 10 data-points). What can be noticed is that the constant 400 is well fitting for couples of players around 2400 rating points.

It seems, thus, that the Elo rating is not a linear scale; the spacing between subsequent values is not the same because the same difference around rating 2400 gives lower advantage than around 1900 (note that the greater the Elo constant is, the lower the advantage of the player that has higher rating).

Let us clarify it further on. Let us assume that we want to keep the Elo model in which the expected score depends merely of the difference of ratings. However, the results of the estimation of the logistic model shows that for the players around 2400, 400 points of advantage gives the expected score about 0.91 (as $1/(1 + 10^{-400/400}) \approx 0.91$)); for lower ratings (around 1800), the same difference in ratings gives the expected score only about 0.76 (as $1/(1 + 10^{-400/800}) \approx 0.76$). That means that the same numerical difference refers to smaller difference in the real strength of players. Thus, if we wanted to keep a unique model with the constant 400, also for weaker players, we could rescale the rating (Fig. 5).

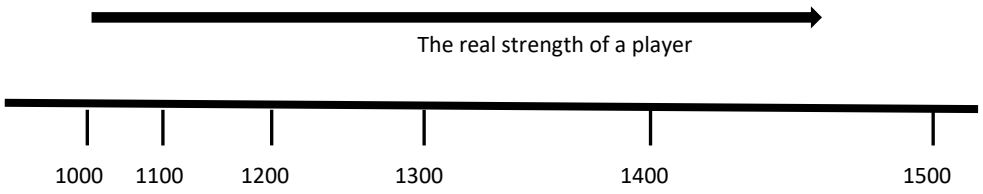


Figure 5. Differences in numerical values not linearly corresponding to the real strength of players. The values and spacing in the figure are given for demonstrative purposes only

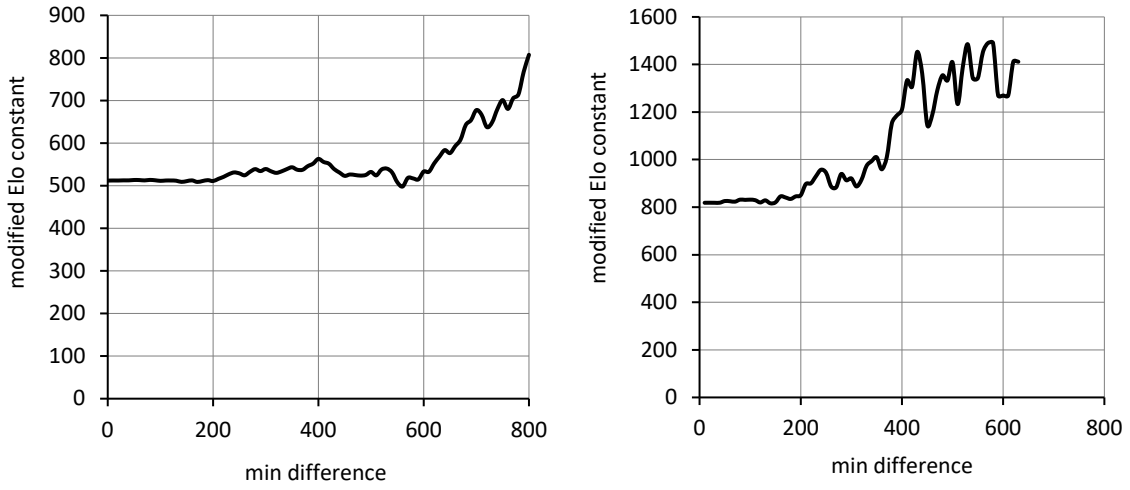
Source: own elaboration.

There is a very interesting lack of monotonicity around value 2000 till about 2050. Does it have anything in common with the fact, that having rating between 2000 and 2200 is considered to advances a player into ‘candidate master’ class? This question will be left open.

3. Underestimating a Weaker Rival

It is not easy to find the real distances between strengths of players as their performance may be influenced not only by their objective strength and random fluctuations. Two more phenomena are very likely to appear, and they have already been described in chess journals. The first one reveals that high-rated players tend to underestimate low-rated players, thus they do not engage fully in the play, which may, supposedly, lead to unexpectedly low results achieved by those better players.

Let us investigate the results of the couples of players differing in ratings with reference to at least some given value (i.e. min. difference in figure). For the increasing difference, we use as input data smaller and smaller sets of games. By estimating the logistic model, we can find modified Elo constants for those limited sets, see Fig. 6a.



a b

Figure 6. Modified Elo constant vs minimal difference of ratings between players (a) The same but restricted to couples of players with both ratings below 2000 (b)

Source: own elaboration.

Indeed, it can be noticed that if the ratings of players differ more than 600, the constant increases rapidly as compared to 512 (which is a constant for the whole set of observations). One explanation is that – as the spacing between lower-rated players seem to be shorter than the spacing between higher-rated players (which was presented in Fig. 5) – it also results in shorter spacing between 'low-rated' and 'high-rated' (as the spacing between these two groups consists of spacing between the lower-rated plus the spacing between the higher-rated). This explanation would not be convincing if a higher constant for players with very different rankings would also occur if they both belonged to low-rated players (a big difference can occur for players below, e.g. 2000 Elo points, as the lowest possible ranking is 1000). After having checked if our data cover such cases, we came to the conclusion that, indeed, the results presented in Fig. 6b suggest that there is the effect of underestimating the weaker opponent (Markos, 2022), and unexpectedly lower results of high-rated player playing with low-rated player is not merely the effect of underrating low-rated players (as it is visible also within those low-rated players). Still, it is only a suggestion, as results presented in Fig. 6b are based on not so large a set of data.

4. The Problem of Short Draws

The second situation that may influence the scores of players is a problem of short draws. It concerns players with similar rankings who prefer to stay with a safe draw than to risk and fight for the win, even if the chances are high. There exists a ranking of most and least aggressive chess champions based on the data collected by Johannes Fischer from tournaments and match games played until 2003 (Fischer, 2004). It turns out that least likely to draw was Steinitz (only 21% of draws) with average length of drawn games equal to 42 moves. On the other hand, Kramnik drew in 57% of games, with the average length of these drawn games equal to 31.

The Elo model does not differentiate between two situations: two draws and, on the other hand, one win and one loss. In both cases, the average score is ½. Still, for the spectators and sponsors, there is a huge difference if the game is interesting and emotional or if it ends up after 15 moves with shaking hands and position completely not clear to be even. Such propensity of players may result in at a bit lower result than possible.

Certainly, statistical data concerning only the final results cannot give any hint if the draw was really forced by the situation (e.g. not enough material to mate of neither site) or was it the 'shake hands' short draw. It is suggested that the draw is a 'natural result' in chess for skilful players, and if neither of them makes a mistake, the result would always be a draw. If that is the reason of frequent draws, we should expect more draws with players of similar strengths, i.e. the stronger both players are, the more frequent draws there should be as it requires a strong player not to make any mistake.

First, we have investigated the probability of a draw for players with a given difference in ratings, without taking into account their absolute strength. The results, for the whole set of players, are presented in Fig. 7.

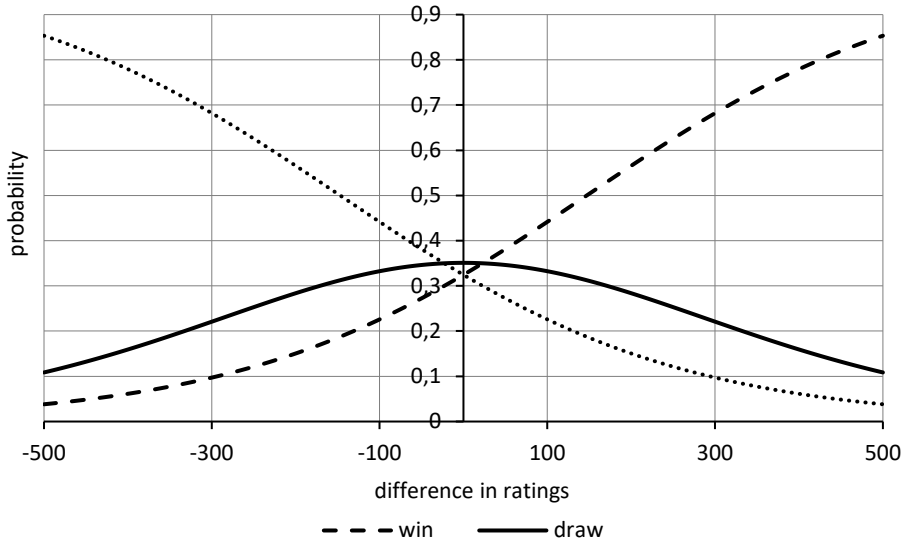


Figure 7. Probabilities of win, loss and draw while having a given advantage over the opponent. Results based on estimation of logistic models for empirical data

Source: own elaboration.

Of course, the sum of all three values for any difference between players is always equal to 1. Obviously, with the increasing difference between players, the probability of a draw decreases and increases the probability that the higher-rated player wins; it is nothing surprising. More interesting results can be observed if we take into account not the whole set of players, but of

players with particular ratings. Figure 8 presents the probabilities of draws for players with three different ratings (1500, 2000, 2500), playing with rivals that have either lower rating ($d < 0$) or higher rating ($d > 0$). Note that this time, the plots are not symmetrical. For the whole set of players, as in Fig. 7, it has to be symmetrical because if a randomly chosen player having an advantage of d points has *p*% of a draw, his or her rival (having a disadvantage of the same strength, d, that is ($-d$)) has to have the same probability of a draw. However, if we choose particular rating of a player, let's say 2000, the probability that she/he will draw with a player who has the advantage of d point does not have to be the same as the probability of a draw with the player who has a disadvantage of d point. In the first case, the rival of our player rated 2000 has a rating of 2000 + d, while in the second case, the rival has a rating of 2000 – d. Thus, if the probability of a draw for the player with a given Rtg and advantage d_1 is equal to p_1, the same probability of a draw has to appear for the player with (Rtg – d_1) and a disadvantage d_1, that is, – d_1 on the horizontal axis; similarly for the player with a given Rtg and a disadvantage d_2.

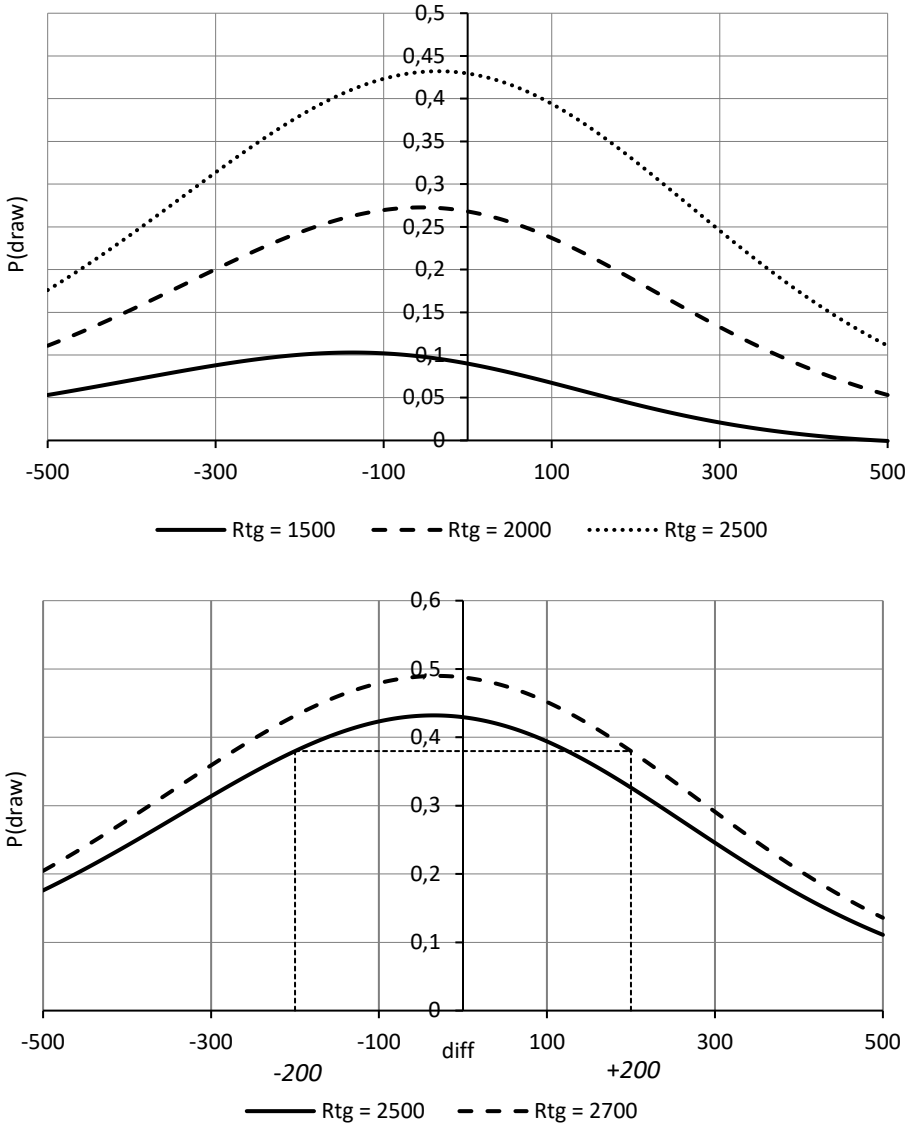


Figure 8. Probabilities of draws for players with ratings 1500, 2000 and 2500, having a given advantage (diff) over the rival (a). Illustration of why the plots in part (a) are not symmetrical (b)

Source: own elaboration.

What may be observed here? According to the hypothesis that the draw is a natural result for good players, the relative frequency of draws increases with the increasing rating of players who do not differ in strength (the increasing value of the intersection of the curve with the vertical axis, i.e. diff = 0). Less understandable is why stronger players are still more prone to draw even if they have an advantage over their rivals. Having 500 points of an advantage, a weak player (Rtg = 1500) will fight to win; however, with the same advantage, his or her stronger equivalent (Rtg = 2500) will draw in more than 10% of cases. Note that this effect runs in the opposite direction to the effect of nonlinearity of scale suggested above. It should be remembered that greater Elo constant in the lower range of scale (as compared to the constant in the higher range of scales) suggests that the spacing between subsequent points in the lower range is smaller than in the upper range. This means that the difference in real strength for better players with the same differences in ratings should be greater; thus, the draws – if not taking into account the 'natural tendency to draw' – would be less likely.

5. The Advantage of Playing White

The last issue related to the Elo model that can be investigated based on the current data is the advantage of playing white (Berg, 2020; Brams & Ismail, 2021). As mentioned above, it is usually expected that playing white is a real advantage and that it is the reason why during matches each player usually plays white as many times as his/her opponent. In Double Round Robin Tournament, like in the 2022 Candidates Tournament in Madrid, each pair of players plays twice so as to change the white player. To explore this question, we have first examined the logistic model for wins and for losses (Tab. 2). So, does the probability of such results depend on playing white? We assume that playing white increases the probability of the win while it decreases the probability of the loss.

Table 2. Results of estimation of models for winning and for loosing

Model for winning				
	estimate	standard error	*z*-statistics	*p*-value
constant	0.72808	0.22711	3.20593	0.00135
diff	0.00535	0.0001	37.09340	3.59216*10^–301
Rtg	–0.00075	0.00010	–7.48434	7.19054*10^–14
col	0.42562	0.05899	7.21553	5.37234*10^–13
Model for loosing				
	estimate	standard error	*z*-statistics	*p*-value
constant	1.15370	0.22799	5.06040	4.18386*10^–7
diff	–0.00460	0.00014	–31.83820	1.91929*10^–222
Rtg	–0.00075	0.00010	–7.48434	7.19054*10^–14
col	–0.42562	0.05899	–7.21553	5.37234*10^–13

Source: own elaboration.

Note that while estimating the model, we have controlled not only for the difference, but also for the ranking of the players. As during the Olympics each couple of players played only once, it might have happened that better players, just by pure luck, were more frequently given the opportunity of playing white. Since we control for the rating of the player and the difference between the rivals, the model parameter determines the effect of playing white ceteris paribus. So, the influence of playing white is isolated, as we compare two situations in which the same player plays with the same opponent, and the only difference between situations is, if the investigated player is playing white or black.

What we can see here is that, indeed, playing white significantly increases the probability of winning and decreases the probability of losing (let us not forget that there is still the third possibility, i.e. drawing). Detailed analysis shows that the better player, the greater advantage for him or her of playing white. In Figure 9, there is a plot of ratio of probability of winning of a player with the given rating (horizontal axis) playing with the player of the same rating; the player playing white (numerator) to playing black (denominator). If this ratio was equal to 1, it would mean that the probability of winning does not depend on the colour. However, it is greater than 1, and the higher the rating is, the greater the ratio is.

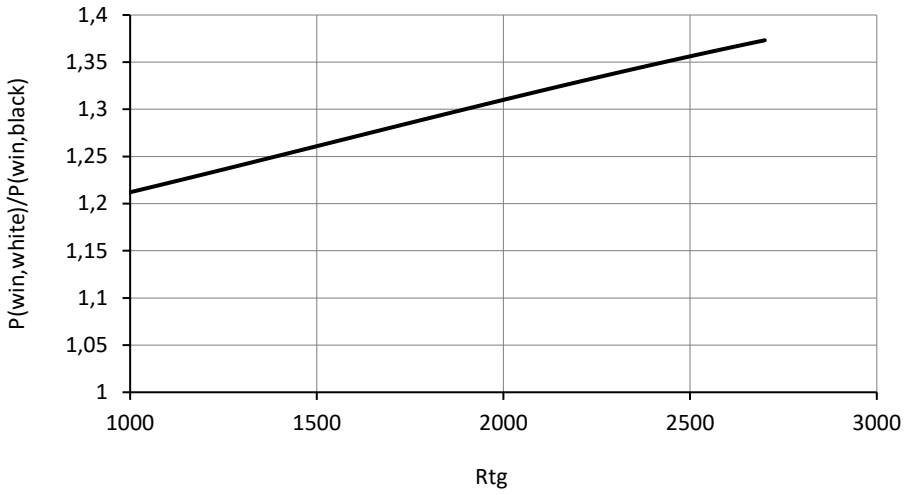


Figure 9. The advantage that gets the player of a given rating when playing white

Source: own elaboration.

That means that playing white always makes it more probable to win; however, the stronger the player is, the greater advantage she/he can make of playing white.

6. Summary

In conclusion, we can say that the Elo model with the constant 400 is suitable only in a very narrow range of ratings of players. In general, this constant seems to be greater than 400 and rather approaches 500. Thus, either we modify the constant of the model or we change the method of updating ratings so as to make equal the differences between the strengths of players with equal differences in ratings, which is the assumption of the Elo model.

Furthermore, we can conclude that strong players tend to underestimate their weaker opponents, thus not winning as frequently as they seemingly could. Additionally, strong players tend not to force themselves playing draws more frequently that their weaker colleagues.

Finally, the Elo model does not take into account playing white or black. Playing white increases significantly the expected score of a player; and the better the player is, the higher the increase. For grandmasters, this increase is even about 35% higher (as compared to the probability of winning playing black). This means, for example, that 2500–Elo–point player has 24% of winning playing black with the opponent with the same ranking (and 33% of losing), but 33% of winning while playing white (and, correspondingly, 24% of losing) – which makes 9 percentage points of difference.

It seems that although the Elo model is the very best one, maybe it requires a bit of corrections.

References

Berg, A. (2020). Statistical Analysis of the Elo Rating System in Chess. *Chance*, *33*(3), 31-38. https://doi.org/10.1080/09332480.2020.1820249

Brams, S. J., & Ismail, M. S. (2021). Fairer Chess: A Reversal of Two Opening Moves in Chess Creates Balance Between White and Black. In *2021 IEEE Conference on Games (CoG)* (pp. 1-4). IEEE.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Publishing.

Fischer, J. (2004). *Most and Least Aggressive World Champions*. Way Back Machine. http://web.archive.org/web/20090208091838/http://www.chessbase.com/newsdetail.asp?newsid=2096

Glickman, M. E., & Jones, A. C. (1999). Rating the Chess Rating System. *Chance, 12*, 21-28.

Markos, J., (2022). *The Winning Academy 13: Facing a Weaker Opponent? Avoid These Mistakes!* Fritz 20. https://en.chessbase.com/post/the–winning–academy–13–facing–a–weaker–opponent–avoid–these–mistakes

Sismanis, Y. (2010). *How I Won the "Chess Ratings – Elo vs the Rest of the World" Competition*. arXiv. https://arxiv.org/pdf/1012.4571

Veček, N., Črepinšek, M., Mernik, M., & Hrnčič, D. (2014). A Comparison Between Different Chess Rating Systems for Ranking Evolutionary Algorithms. *Annals of Computer Science and Information Systems*, *2*, 511-518. https://doi.org/10.15439/2014F33

Ocena systemu rankingowego ELO na podstawie wyników 44. Olimpiady Szachowej w Chennai

Streszczenie: Szachiści opisywani są tzw. rankingiem Elo, który ma odzwierciedlać siłę gracza i jest aktualizowany na podstawie jego wyników, czyli zwycięstw, porażek oraz remisów. Wygrana lub remis z graczem o wyższym rankingu powoduje wzrost rankingu danego zawodnika, natomiast porażka lub remis z graczem o niższym rankingu obniża jego pozycję. Zyski i straty w systemie rankingowym są obliczane według określonego algorytmu. Ranking Elo ma także przewidywać oczekiwany wynik zawodnika mającego daną przewagę lub niekorzyść wobec przeciwnika. Pytanie, jak dobre są te przewidywania? W niniejszym artykule badana jest dokładność modelu Elo w oparciu o wyniki 44. Olimpiady Szachowej w Chennai.

Słowa kluczowe: model Elo, ranking szachowy