Bias in the ELO-system of online chess

Keywords: false, ELO, glicko, ranking, UCFS

Introduction

There are millions of chess players around the world that play chess on websites like chess.com. According to chess.com, there are 600 Million chess-players worldwide, and more than 20 Million members on chess.com that play up to 1 Million games per day, and there are 360,000 tournament players and 1594 grandmasters of which only 2.2% are female, in fact, although it is free to play for everyone. Today, bias in research is ubiquitous, also in chess research and rating, which should be analyzed. The strength of a player in the game of chess and additional zero-sum games is usually estimated and assessed in ELO, a rational chess skill rating system that was developed by Arpad Elo, a Hungarian physics professor in the USA. The ELO rating system statistically derives numerical outcome in ELO of the games a player plays against opponents of different strength: it increases the ELO score value of the winner as much as it decreases the ELO score of the loser. A high-rated player can lose more points against a low-rated player, who can win more points against a high-rated player, and vice versa, a high-rated player can win fewer points against a low-rated player, who will lose fewer points against a high-rated one, leading to ELO conservation, like the conservation of energy, in the system. This adjustment upwards or downwards goes back to Elo¹,² who suggested and USCF, the United States Chess Federation, has implemented his suggestions in 1960. ELO’s formula (3) for the logistic probability function that expresses the expected score that is scalable for ELO changes, is shown here:

\[
ExScore_A = \frac{1}{1 + 10^{ELO(A) - ELO(B)/400}} ; \quad ExScore_B = \frac{1}{1 + 10^{ELO(B) - ELO(A)/400}}
\]

Hence, an ELO difference of 400 results in a 10-time scaled expected score.³ If a specific player is expected to win in 40% of all points in games against a stronger player and wins 50% of them, then the ELO system would rise but would stay the same at 40% of wins. The scale of the scoring is given by a factor x.⁴

There have been some slight advancements but mainly in the fine-tuning of the ELO system, such as the Glicko or Glicko–2 ELO system⁵ that is also used on online chess–sites, like chess.com,¹ or by the Australian Chess Federation, and further online chess sites. Instead of factor x, Glicko–1/2 uses the statistical rating deviation to integrate the error of deviation from the ELO values of the two players.

Results and discussion

Millions of professional and hobby chess players are playing and learning chess on online chess gaming sites. Online chess sites like chess.com¹ play also a big role on chess of today, the chess community, the chess culture, but also chess learning and training and more than 100 Million players have used and played on an online chess site. These sites provide great benefits for chess players as they concentrate information, tools, news, articles, statistics, games, analysis, tools, opening books, lessons, ideas, puzzles, and much more to chess players. They represent an ideal place to quickly find an opponent of your strength and length of the game you want to play. Since a huge proportion of chess players use these sites, chess–websites turns into a new infrastructure of chess at some point that must function correctly. It is a transition from a private function to a public platform and infrastructure that many websites should go through today, whenever a societal function is in play. Here, it is very important that online chess sites have a functioning, reliable and comparable ELO–system in place, as the success in chess for most players is found in ELO values, especially for the majority of players that visits these sites. The initial and longstanding idea of ELO’s chess rating is a better comparability of chess players strength, which Dr. Elo⁴ achieved, and Glicko or Glicko–2 can be seen as a statistical refinement, a fine–tuning that makes use of given rating deviations and players to achieve a faster real ELO adjustment with less volatility and games. Usually, it starts at 1500 and a rating deviation of 350, which is 300 ELO higher than chess.com,¹ where everybody has to start with only 1200 ELO points. Glicko is a rational and suitable amendment to the ELO system that initially manually used a factor x for scaling. Glicko ELO–rating uses the rating deviation instead to assure a higher certainty and a lower volatility.

As a result, these novelties should only further improve the speed of accuracy in ELO by fewer games and the deviation of ELO ratings, which both makes the ELO system more accurate and comparable, if done right. However, since 10–20 years, many players might have noticed it; it has become seemingly impossible to obtain a comparable ELO-rating on one of these chess sites in the range of 1000–2000 ELO. The question arises how such a widely “perceived bias” can or could happen in such sophisticated and statistically elaborated ELO or the Glicko–1/2 rating system that stem from Harvard²,³–⁵ It is clear that any ELO rating only makes sense if it is comparable among all platforms and players. It is the idea of any chess rating system that is must be comparable. You cannot say we are using Glicko and thus you cannot compare us with another rating system – they all must be comparable, Glicko is just the refinement of ELO, only a fine tuning if you will. So, these questions arise: are they really not comparable, what is the deviation to real ELO values, and how can you make them comparable again?
The answer to these questions could be given by the following hypothesis about the rating: Whenever a new player enters a site like chess.com or Chess base, Chess Cube, ICC, he or she will start with a number of ELO points and a rating deviation in the case of Glicko/Glicko-2. These sites have always argued that this would not be a problem as over time the player will receive ELO value, and Glicko would even accelerate this by using more suitable rating deviations. But like in the mathematical field of differential equations, the initial value problem arises. The 'initial condition' would be the starting ELO value and its rating deviation in a system of entities that have different strength and likelihoods to win, draw and lose. As a result, the amount of ELO-points will stay at a specific aggregated level in the system, which can be too low for all players. Let's 'hypnotize', there are three players in the game: one player with real skill strength (RSS) of 1000 ELORSS, one with 2000 ELORSS, and also one with 3000 ELORSS. If these players would start with 1200 ELO points each – as on chess.com – and would now play many games, it would result in a fair rating only for the 3000ELORSS player, who could maybe achieve 2900 ELO on the chess-web-site. But the player with an RSS of 2000 ELO could only achieve 1600 of the remaining ELO points in the system, and the player with 1000 ELO could receive 0 ELO points this way on this website for the three players. This hypothetical example game illustrates what kind of issue the ELO system generates that maintains a constant amount of ELO points per system. Grandmasters at a level of 3000 might be still a tiny minority, and there are worldwide only 4–5 players with a real rating of 2900–3000 ELO points. The majority of players will still have an ELORSS between 1000 and 2000 ELO points, which is the part of the statistical distribution that would be the most erroneous and false rating due to the artificial scarcity of overall ELO points in the entire system – caused by the Cauchy problem, the too low 'initial value' of 1200 ELO points or 1300, or 1400, depending on the chess-site.

To test this hypothesis, the number of players in blitz games was compared between official RBB FIDE ratings and blitz games on chess.com by plotting the relative amount of players per 100–ELO-segment (Figure 1). An unexpected gigantic shift in the curve and peak can be, in fact, observed (Fig. 1). Almost 1000 ELO points subdivide the peaks of both ELO rating systems (Fig. 1). Hence, this further indicates that comparability seems most likely not to be given between such FIDE–ELO and chess.com ELO-ratings systems. Comparability of ELO system seems not to be given for tournament ELO and other ELO ratings.

To solve this problem, every chess ELO rating system must advise the right ELO points to players: this means, a 1000 ELORSS should achieve 1000 ELO over time, a 2000 ELORSS player should receive 2000 ELO points and a 3000 ELORSS player must still get 3000 ELO. The challenge is to identify the ELO reflecting the real skill strength (RSS) of the individual player. If the initial value, the starting ELO, would be correct, the problem would be much smaller as the overall amount of points to compete for is more correct. But the players would still get better over time, one might assume, and this would not increase the overall amount of ELO points in the system – it is a monetary-type ELO problem. Centrally, one should not cause ELO inflation nor a deflation; one should assure stability and comparability so that ELO values are correct for all players, which still drift over time as players come and go, get worse or better. To do this, the overall aggregated ELO points must reflect the overall ELORSS of all the players.

One can also not assume that the majority of all players have ELO strength of 800 (Figure 1), which approximates a stage of learning the major rules of chess. There are too many chess players to assume this. People that visit chess.com are chess fans and people interested in chess, they are really not so bad players one might surely assume, also in light of up to 600 million chess–players worldwide.

One way to technically do this is shown by Elometer that uses item response theory to derive an ELO maximum–likelihood estimate using a set of chess problems of “the Amsterdam Chess Test” with known properties, based on a prediction formula regressing the “RSS–skills” of 259 players, all participants of the Dutch Open Tournament, ranked from 1169 to 2629 on the Birnbaum model.

Elometer thereby tries to reveal the real ELO strength of a player using the above mention methods. By using Elometer, the hypothesis that the ELO system is erroneously biased, and not comparable, could be again further substantiated: Atypical chess.com player with a skill strength of 1900 ELO points in puzzles, 2000 in lessons/position strategy, achieved an ELO of 2015 at Elometer (95 CI, 1891–2140), which is shown in Figure 2 and compared to the chess.com ELO rating (Figure 3), which is much lower at 1475 ELO. Hence, two things can be derived: (i) there is no ELO rating comparability but an urgent requirement of it and (ii) Elometer provides a key to find a first solution: by assessing the RSS–ELO.

To test this hypothesis, the number of players in blitz games was compared between official RBB FIDE ratings and blitz games on chess.com by plotting the relative amount of players per 100–ELO-segment (Figure 1). An unexpected gigantic shift in the curve and peak can be, in fact, observed (Fig. 1). Almost 1000 ELO points subdivide the peaks of both ELO rating systems (Fig. 1). Hence, this further indicates that comparability seems most likely not to be given between such FIDE–ELO and chess.com ELO-ratings systems. Comparability of ELO system seems not to be given for tournament ELO and other ELO ratings.

To solve this problem, every chess ELO rating system must advise the right ELO points to players: this means, a 1000 ELORSS should achieve 1000 ELO over time, a 2000 ELORSS player should receive 2000 ELO points and a 3000 ELORSS player must still get 3000 ELO. The challenge is to identify the ELO reflecting the real skill strength (RSS) of the individual player. If the initial value, the starting ELO, would be correct, the problem would be much smaller as the overall amount of points to compete for is more correct. But the players would still get better over time, one might assume, and this would not increase the overall amount of ELO points in the system – it is a monetary-type ELO problem. Centrally, one should not cause ELO inflation nor a deflation; one should assure stability and comparability so that ELO values are correct for all players, which still drift over time as players come and go, get worse or better. To do this, the overall aggregated ELO points must reflect the overall ELORSS of all the players.

One can also not assume that the majority of all players have ELO strength of 800 (Figure 1), which approximates a stage of learning the major rules of chess. There are too many chess players to assume this. People that visit chess.com are chess fans and people interested in chess, they are really not so bad players one might surely assume, also in light of up to 600 million chess–players worldwide.

One way to technically do this is shown by Elometer that uses item response theory to derive an ELO maximum–likelihood estimate using a set of chess problems of “the Amsterdam Chess Test” with known properties, based on a prediction formula regressing the “RSS–skills” of 259 players, all participants of the Dutch Open Tournament, ranked from 1169 to 2629 on the Birnbaum model.

Elometer thereby tries to reveal the real ELO strength of a player using the above mention methods. By using Elometer, the hypothesis that the ELO system is erroneously biased, and not comparable, could be again further substantiated: Atypical chess.com player with a skill strength of 1900 ELO points in puzzles, 2000 in lessons/position strategy, achieved an ELO of 2015 at Elometer (95 CI, 1891–2140), which is shown in Figure 2 and compared to the chess.com ELO rating (Figure 3), which is much lower at 1475 ELO. Hence, two things can be derived: (i) there is no ELO rating comparability but an urgent requirement of it and (ii) Elometer provides a key to find a first solution: by assessing the RSS–ELO.

![Figure 1](image1.png) Comparison of two Statistical Distribution of Rated Players in Blitz Games at Chess.Com (blue) and FIDE (orange). Please notice the shift of 1000 ELO points of the peaks in the two rating systems.

![Figure 2](image2.png) Scientific ELO estimation based on Elometer's item response theory (January 2018). A 95% confidence interval (dashed lines) is shown for the estimate (solid line). Please notice, this tournament-related scientific estimate of 2015 ELO corresponds to only 1475 ELO at chess.com in the example.

Bias in the ELO system of online chess

If all new chess players would do a testing like the Elometer assay\(^7\) this could improve the initial problem, the starting problem, of all players that enter the gaming system. Consequentially, the total amount of ELO points in the system would be more accurate and the slope and steepness of the curve would not be so much distorted as it is today (Figure 1). One could assume, that there are only a few players that start at the right ELO, mainly grandmasters that enter the site and may start at their known ELO value, but the majority of players are not grandmasters and 20 Million false initial valued ELOs dramatically shift the entire ELO curve and would make it extremely steep—as can be seen in Figure 1.

Another way to solve this is to always release enough ELO points into the competitive situation of the chess site. Simply calculate what one has to expect in a normal distribution of chess hobby players and professionals, take the overall points, and fill up the online chess game site with additional ELO points up to the level of the aggregated sum of all of these ELO values—so that enough points are in the system that represents a normal distribution of chess players. But how to best assess the normal distribution of all players ELO\(_{RSS}\)? Estimates can be based on (A) historic tables or (B) ELO\(_{RSS}\) tests.\(^7,9\)

Hence, this paper concludes that the ELO rating must be always comparable at any given level of ELO and play. The individual ELO score can only be correct if the overall aggregated ELO points in the system are correct too. This can only be achieved with right initial ELO scores and by introducing new freely floating ELO points into the entire system up to the level of the expected aggregated normal distribution, and subsequently letting all of the players compete for them in a fair way without ELO scarcity caused steepness of the ELO slope between from 100 to 2000 ELO. Falseness and bias seems not to be an issue for the grandmasters on the site as they enter the site with their own points and they are maintained in a bubble and only play with grandmasters that brought their own points—there is no much contact with a real chess player’s world where you face harsh ELO-slopes and systemic ELO scarcity. Elo inflation and deflation must be avoided as much as artificial scarcity of ELO points and steepening of the ELO slope. Millions of frustrated chess players might want to have their reliable ELO.

Acknowledgements
None.

Conflicts of interest
The authors declared that there are no conflicts of interest.

References
1. www.chess.com