Bias in the ELO-System of Online Chess

Abstract

ELO is the key performance indicator in chess, a global and worldwide measure of chess skills and strength of chess play. Since their invention, ELO-systems are intended to be comparable and unbiased, so that chess players can know their level of play and can compare it among systems and internationally. Moreover, ELO is the defining feature of grandmasters with legal implications. Thus, it is highly important to only have unbiased ELO-system, whenever a terminology bearing ELO is used. However, most recent online chess sites seem to utilize highly biased ELO-system that can maybe not be compared to tournament or FIDE chess ELO-score systems. It could be that online-ELO-systems are strategically biased to make all hobby players around the world appear less professional by systemically down-shifting the ELO-score system. This short report offers some insights about this potential ELO-deflation phenomenon in online chess. It can be shown that online ELO systems do not reflect real chess skill strength in ELO as measured by different methods: (1) game strength does not reflect tactical or strategic strength, (2) there is a shift in ELO-distribution between hundreds and up to thousand ELO-points of the average level player, e.g. FIDE RBB versus an online site, (3) Elometer, a scientifically approved and scientifically standardization method also reveal this dramatic discrepancy in ELO-reduction. The reason for this deprivation of ELO-points seems to be artificial ELO-scarcity that is introduced into the system whenever a new player enters the site and has to start with fewer ELO-points than his or her skill level. There are not enough aggregated ELO-points that mirror the sum of all chess skills in online chess and year after years it seems to worsen, while FIDE-ratings are rising slightly. Hence, this work reminds that the total ELO-points per rating system must be always adjusted. Millions of chess-players world-wide might be extremely underrated and their chess skills overlooked.

Keywords: false, ELO, glicko, ranking, rating, chess

Introduction

There are millions of chess players around the world that play chess on websites like chess.com.1 According to chess.com,1 there are 600 Million chess-players worldwide, and more than 20 Million members on chess.com1 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, while it is free to play for everyone.1 Today, bias in research is ubiquitous, also in chess research and likely also in ELO rating, which … rating,2 which should be further analyzed in some more details. 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 initially developed by Arpad Elo, a Hungarian physics who lived, worked and taught in the USA, in the USA. The ELO-rating-system statistically derives numerical outcomes 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, roughly speaking, deviations of the rule equal much out as the sum of both directions. 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 leading to an ELO conservation effect after deviations and correction factors equal much out, slightly comparable to the conservation of energy that is fixed in the system. like the conservation of energy, in the system. This adjustment upwards or downwards goes back to Arpad Elo3,4 who suggested it, and USCF, the United States Chess Federation, implemented his suggestion in 1960, subsequently it became an international standard, and USCF, the United States Chess Federation, has implemented his suggestions in 1960. The basis of the ELO-formula (reference 3) for the logistic probability function… ELO’s formula style as reference 3 for the logical probability function expresses the expected score that is always further scalable for ELO changes per game, is represented by the expected score (ExScore) of player A and B:

\[ \text{ExScore}_A = \frac{1}{1+10^{(ELO_B-ELO_A)/400}} \]
\[ \text{ExScore}_B = \frac{1}{1+10^{(ELO_A-ELO_B)/400}} \]

Hence, an ELO difference of 400 results in a 10-times scaled and the expected score is part of the ELO-change calculation (3,4ref); if a specific player… is expected to win 40% of all points in games against a stronger player and wins 50% of them, then the ELO would rise but would stay the same at 40% of wins, in general. The scale of the scoring is given …by a factor x, which modulates the amount of ELO-point-changes and which generally does not much alter the overall aggregated ELO points in the system. There have been some slight advancements but mainly in the fine-tuning of the ELO-system, such as the Glicko or Glicko-2 ELO-systems that is used on online chess-sites, like chess.com, or by the Australian Chess Federation, and further online chess sites. As a suitable improvement 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. As a result, Glicko systems faster adjust the ELO rating accurately but only up to the level of the specific rating system and the historical preset ranking, which can be either correct or not, right or wrong, correct or giving a too high or too low ELO scoring.

Roman Anton
Department of Theoretical Sciences, The University of Truth and Common Sense, Germany

Correspondence: roman.anton@gmail.com

Received: January 08, 2018 | Published: March 09, 2018
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.com 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 are using, mainly or only, chess websites to play, these sites are becoming a commons infrastructure and platform of chess, like any other major websites, like search engines. They are already today a commons for which all users get a right of correct treatment, such as the ELO-comparability, which can be seen as a qualification like a certification or an official testing of a skill, which of course must be unbiased and correct. No matter if Glicko-1, Glicko-2, or UCFS, all ELO-rating systems must be always comparable to one ELO-scale, per game type, e.g. rapid or blitz, by using a simple standardization. Some point must function correctly. It is a transition from a private function to a public platform and infrastructure that many websites are going through today, whenever a societal function is in play, the user rights must be better protected. 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 the ELO-values, especially for the majority of players that visits these sites. The initial and longstanding idea of Apard Elo’s chess rating is a better comparability of chess players strength, which Dr. Elo had achieved, while Glicko or Glicko–2 can be seen as a statistical refinement, a fine-tuning that makes uses of given rating deviations and players to achieve a faster real ELO-adjustment with less volatility and volatility and less games needed - but it does thereby only partially correct the overall ELO-points in the system, which still leads to a ELO-deflation over time if average skill strength is higher than average ELO-starting value and vice versa. Initially it was configured to let players start with 1500 ELO-points and a rating deviation of 350, which is 300 ELO-points ... higher than on chess.com, where everybody has to start with only 1200 ELO points. Thus, the chess-sites broke the most basic standardization of a 1500-ELO starting value. The average skills of players is even higher and not lower, one must assume as an ELO of 1600 on a scale of 3000 already represents a player that can beat 999 of 1000 chess players worldwide, as the top 200,000 and top 0.1% of all players. Glicko is a rational and suitable amendment to the ELO system that initially only used a manually adjustable a factor x for the scaling that could differ between top players and amateur tournaments or matches. Glicko ELO-rating uses the rating deviation instead to assure a higher certainty and a lower volatility in a faster way.

As a result, these novelties further improved the speed of accuracy in ELO in 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 should bear the quality of Harvard. 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 it 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 inbetween rating systems and also worldwide?

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 ChessBase, Chess Cube, ICC, he or she will start with a number of ELO points and a rating deviation in the case of Glicko-1 and Glicko-2. These sites have always argued that this would not be a problem as over time the player will receive is 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 If we game-theoretically assume, there are the three following players in a hypothetical tournament, there are three players in a hypothetical tournament: one player with a real skill strength (RSS) of 1000 RSS-ELO, one with 2000 RSS-ELO, and also one with 3000 RSS-ELO. If these players now all 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 3000 RSS-ELO player, who could maybe achieve 2900 ELO on the online chess website, as nobody has the RSS to deprive his ELO points. 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 player, as 1-2 players can deprive their points. Glicko could only slightly improve this. This hypothetical example game illustrates what kind of issue the ELO system generates that maintains a too 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 RSS-ELO 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 online chess website.

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). In fact, an unexpected gigantic shift in the curve and peak can be 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-rating-system. Comparability of ELO-rating-systems seems not to be given for tournament ELO and other ELO ratings. A major deflation of ELO points must be assumed on online chess websites, as all of them have too low initial ELO values.

To solve this problem, every chess ELO-rating-system must from the beginning advise the right ELO-points to every new player: this means, a 1000 RSS-ELO player should achieve 1000 ELO.
Bias in the ELO-system of online chess

over time, a 2000 RSS-ELO player should receive 2000 ELO points and a 3000 RSS-ELO 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 RSS-ELO 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 good chess players in the world 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. If only 0.33% of players reach a level of 1600 ELO there must be something wrong due to an huge ELO-deflation.

One way to technically do this is shown by Elometer that uses item response theory\textsuperscript{7,8} to derive an ELO maximum–likelihood estimate using a set of chess problems of “the Amsterdam Chess Tests\textsuperscript{7} with known properties, based on a prediction formula regressing the “RSS” of 259 players, all participants of the Dutch Open Tournament, ranked from 1169 to 2629 on the Birnbaum model.\textsuperscript{1,10}

Elometer\textsuperscript{7} thereby tries to reveal the real ELO-strength of a player using the above mention methods. By using Elometer,\textsuperscript{7} the hypothesis that the ELO-system is erroneously biased, and not comparable, could be again further substantiated (Fig. 2): A typical 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\textsuperscript{7} ELO rating (Figure 3), which is much lower at 1475 - indicating a potential deflation of about 500 ELO-points in online chess. Hence, two things can be derived: (i) there is no ELO-rating comparability but an urgent requirement of it and (ii) Elometer\textsuperscript{7} provides a key to find a first solution: by assessing the RSS-ELO that could be theoretically used for initial values.

If all new chess players would do a testing like the scientific Elometer assay\textsuperscript{7} this ELO-number could be used to improve the initial rating problem that causes the deflation, of all players that enter the gaming system. The learning curve of all players could be also assessed in this way to adequately increase the overall amount of ELO-points. Consequently, 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 who may start somewhere around or at their known ELO-value, but the majority of players are not grandmasters and 20 Million false initial valued ELO-scores could, in fact, dramatically shift the entire ELO-curve and would make it extremely steep–as can be maybe seen in Figure 1, the distribution is already falling after 1000, while it should be still rising according to the estimates, and the RBB rating is also rising until 1800 - a difference of 800 ELO-points (Fig. 1), the pools of players are a bit different but the deflation remains likely. But how big and strong is the ELO-deflation?

Another way to solve this is to always release enough ELO-points into the competitive situation on the chess site. Simply calculate what

Figure 1 Comparison of two statistical distributions of rated players in blitz games at chess.com (orange) and FIDE RBB (blue) in 2018. Please notice the dramatic shift of 1000 ELO-points of the peaks of the two systems of rating.

Figure 2 Scientific ELO estimation based on Elometer’s item response theory (January 2018) as a further proof of the ELO-deflation hypothesis. A 95% confidence interval (dashed lines) is shown for estimate of the mean ELO (solid line). Please notice, this tournament-related scientific estimate of 2015 ELO-points corresponds to only 1475 ELO-points on the online chess site chess.com in this example - a striking difference of 540 ELO points revealing a major ELO deflation in online chess ratings especially for all players with an ELO-RSS of 1000-2000.

Figure 3 Please notice, this tournament-related scientific estimate of 2015 ELO corresponds to only 1475 ELO at chess.com in a representative example. An ELO of 1475 recently relates to the upper 10% (percentile 91.1%) of the players on this site in January 2018, this seems to be very biased and atypical for real-world chess and a curiosity of online chess only–that should be fixed: via right initial ELO values and by assuring a suitable amount of overall ELO-points at all times and for all players.

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 all that only minimally change over time due to the vast amount of players, 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

Conclusions

Hence, this short report concludes that an ELO-deflation is likely to be found in most online chess rating-systems as most if not all use too low initial ELO-values. Consequently, all online ELO-rating systems should be checked and made comparable for all players, whenever a deflation (or inflation) can be found. The first Elo-Systems circumvented this problem by giving adequate initial ELO values, this is also done in chess clubs around the world. Reasonably, this should be also done in online chess to avoid an ELO-deflation of a bigger 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 fair way. in a fair way This should be done without any artificial ELO-scarcity that could cause a steepness of the ELO-distribution especially between 100-2100 ELO, what could be maybe called an ELO-deflation shift whenever it is verified. The magnitude of the deflation seems to vary from ELO-level to ELO-level and one can only speculate about how many hundreds of ELO-points are deviating from a player’s RSS-ELO. An ELO-gold standard is missing and should be included for all levels of chess players to prevent any deflation that happens over time. Glicko only reduces the deflation but does not fully prevent it as it benchmarks with the aggregated points and very slightly but still significantly reduces them every new player. Hence, an optimization of Glicko is here suggested: (a) correct initial ELO starting values and (b) any points lost against a new player must be added back to the system until his natural ELO is reached, (c) learning curve effects of the pool of players could be included in times of improved learning, (d) overall aggregated ELO-points must reflect overall aggregated RSS-ELO points. ELO biases 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 top players bubble awhere they might mainly play with other grandmaster-level players. There is not so very much contact with the average chess player’s world where you face harsh ELO-slopes and systemic ELO-scarcity, as there are only very few players better than them that would be able to deprive them of the artificially scarce ELO points. As a result, mainly the stronger players could cause this ELO-deflation shift for weaker players, or the entire curve could be shifted also. Generally speaking, ELO-inflation and ELO-deflation must be better avoided in online chess and this requires active measures that refill the pool of aggregated ELO-points up to the level of the recent ELO-RSS of all players. ELO-gold-standards and ELO-tests like the Elo-meter assay could be utilized to adjust the initial ELO-values, or to better normalize the ELO must be The intrinsic trend towards artificial scarcity of ELO-points and ELO-score aggravation steepening from 100-2000 ELO points must be ended via suitable normalizations and enough total ELO points in the system. Millions of... scarcity of ELO points and steepening of the ELO slope. Millions of frustrated chess players might want to have a more accurate, a more reliable and amore comparable ELO-rating - and millions of player might have been systemically underrated since decades. ELO is a certification like IQ or an exam grade that must meet the claim to be correct. Millions of players might have been damaged in sum by this systematic ELO-downgrade for the majority of hobby-players: for instance, on the job market, in the eyes of others, and even in their own eyes. Most players could be much better than previously anticipated. This should be tested further and steadily for all online chess websites to keep all ELO rating systems comparable.

Acknowledgements

University of Truth and Common Sense is acknowledged here.

Conflicts of interest

The authors declared that there are no conflicts of interest.

References

1. www.chess.com