The Glicko system for beginners

By Michalis Kaloumenos

Member of FIDE Qualification Commission

Reference

The Glicko system was introduced by Prof. Mark Glickman in 1995. The equations used for rating calculations can be found here: http://www.glicko.net/glicko/glicko.pdf

The Clicko-2 rating system, an improvement of the original Glicko system introduced in 2000, was in provisional patent but now is in the public domain. The equations used for rating calculations can be found here: http://www.glicko.net/glicko/glicko2.pdf

There are a lot of papers in Prof. Glickman's personal website (http://www.glicko.net/). Here (http://www.glicko.net/research/acjpaper.pdf) is a 49 page long "Comprehensive guide to chess ratings" although reference to FIDE rating system is out of date since the paper dates back to 1995. This paper (http://www.glicko.net/research/glicko.pdf) also provides detailed mathematical explanation of the Glicko system.

The Australian Chess Federation already uses a slightly modified version of the Glicko-2 system for the calculation of their National rating points. More information can be found in their website: http://www.auschess.org.au/constitution/Ratings_By-Law.txt

This page (http://www.kaggle.com/c/ChessRatings2/forums/t/473/more-about-glicko) from Jeff Sonas' kaggle project provides valuable information for the tuning of the Glicko system.

Reliability of measurement – introduction of "Rating Deviation"

The novelty of the Glicko system is the introduction of a new measure assigned to each chess player together with his/her rating. It is called "rating deviation" (RD) and addresses the problem of reliability of each player's rating. The use of terms such as "reliability" and "accuracy" of measurement becomes clearer with an example from car races:

In F1 racing all teams use telemetry for real time observation of a car's condition, which means that every 2 milliseconds the car transmits to the garage a set of data taken from sensors placed on the car. Suppose that this set of data includes variables such as the position of the car, velocity, direction of movement and the vector of acceleration. From this information it is possible to calculate the position of the car one centisecond (0.01 sec) later with satisfactory accuracy. However those familiar with speed, circuits and race conditions understand that it is almost impossible to calculate the position of a car one second later using this information. Data one second old are too old to provide an accurate measurement of a car's position.

In chess there are many reasons why a player's rating should be considered inaccurate or unreliable, the case of an unrated player (addressed below) being the most extreme. A player inactive for a rating period (or a number of rating periods) is an example. This time of inactivity should not be confused with the reasons that kept the player away from tournament play. Regardless of maternity rest or university exams or intensive study of chess theory, it is irrelevant if the player has improved or worsened during his/her absence. What is important here is that his/her last published rating is old for rating calculations. In addition, the Glicko system considers a player's rating unreliable if it is based on insufficient game results. The relation between number of games and rating deviation is discussed below.

The rating deviation measures the uncertainty in a rating. High RDs correspond to unreliable ratings. A high RD indicates that a player may not be competing frequently or that a player has only competed in a small number of tournament games. A low RD indicates that a player competes frequently. From a statistician's point of view a player rating lies with 95% confidence within the interval:

(Rating − 2 x RD, Rating + 2 x RD)

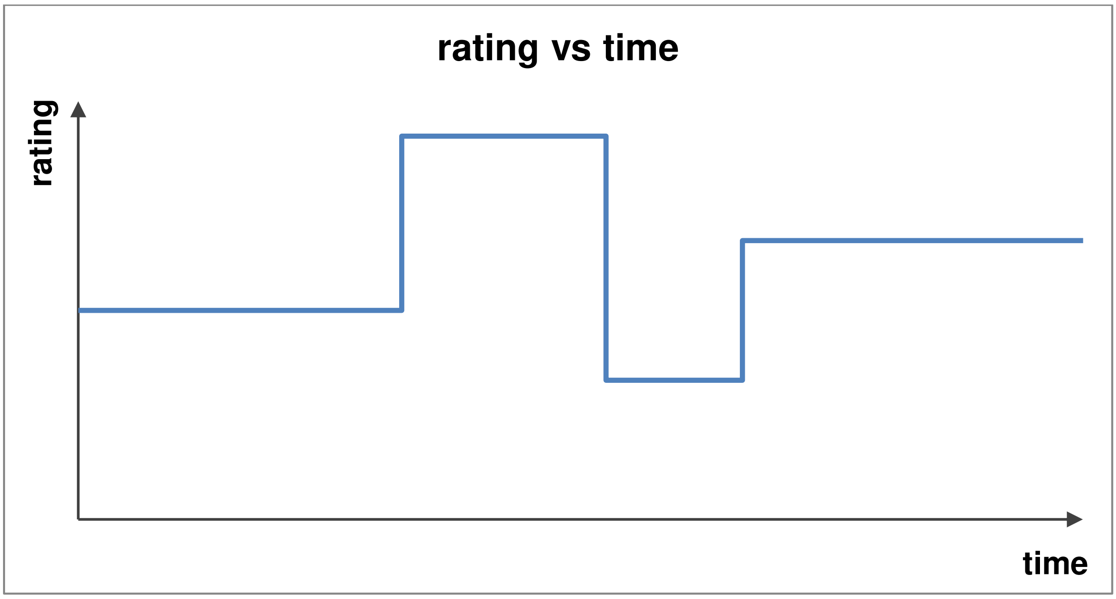
It must be noted here that this confidence interval also exists in ELO rating system. However, the statistical error of ELO calculation is omitted in every day practice. All the same, the middle of the above interval should be published as the player's Glicko rating, but RD must also be published as its value is used for the rating calculation of the next period. However the rating list published by the Australian Chess Federation uses symbols instead of numbers in order to define rating reliability.

The Glicko-2 system introduces a third number for each player called rating volatility (σ). The volatility measure indicates the degree of expected fluctuation in a player's rating. The volatility measure is high when a player has erratic performances (e.g., when the player has had exceptionally strong results after a period of stability), and the volatility measure is low when the player performs at a consistent level.

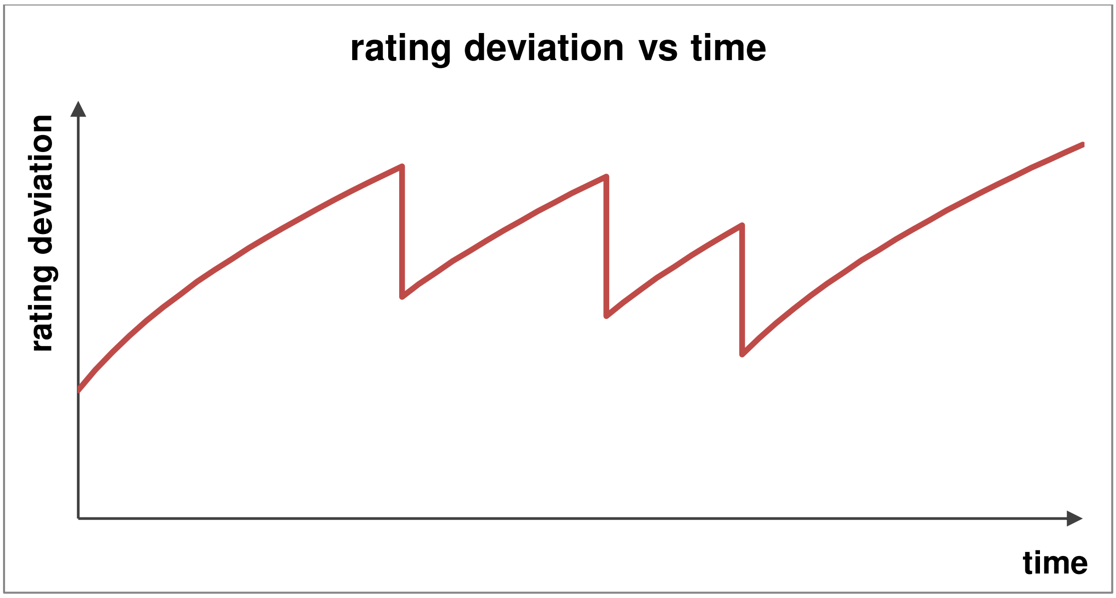
It is obvious that Glicko rating calculations are more complex compared to the ELO system. Today, a copy of the scoring probability table and a pencil are enough for a player to calculate his/her rating change after a game, since all calculations can be easily done manually. The Glicko system requires (for example) a properly prepared Excel file instead. From a historic point of view, Prof. Arpad Elo has deliberately simplified the mathematical equations back in 1960, in order to provide a simple and reliable rating system that can be easily used by every chess player. Nowadays, the use of computers makes this requirement of the rating system not necessary. A possible adoption of the Glicko system will soon make available numerous tools and applications for laptops, notebooks, ipads and cellphones that will help players to input tournament results and calculate rating changes.

The behavior of rating variables vs time

The rating of a chess player is a property that accompanies him/her from the first day that his/her name appears in FIDE rating list and for their entire lifetime. Its graphic representation versus time should be a continuous horizontal line all the time between two games. When a game is over and the scoresheets are signed a new rating is calculated (of course it is possible that it remains the same) and a new horizontal line continues onwards representing the new rating after a single point of discontinuity. The following graph demonstrates this. A vertical line is drawn to connect the two lines in order to emphasize the fact that a chess player cannot lose his/her rating. The graph of rating vs time is the same for both ELO rating system and Glicko. Whenever a player answers the question “what is your rating?” or “what was your rating 4 years ago?” he/she actually refers to a point on this graph:



Rating deviation can also be represented by a continuous line. In contrast to the rating itself, RD rises continuously all the time between the completion of two games following the points of a parabolic curve, then drops vertically when a result occurs and continues upwards again following a new parabolic trajectory. Points of discontinuity when a game is completed are out of the scope of this paper. The following graph depicts this continuous change, however the scale of x- and y-axis are only indicative of the shape of the curve and do not correspond to any particular data.



Prof. Glickman suggests that the value of RD must have a limit superior. Whenever the value of RD is calculated greater than 350, then it must be set to 350 (RD = 350). This value indicates the rating deviation of an unrated player. In most cases, the calculated value of RD can be safely rounded to the nearest integer without causing any significant error in calculations. It is possible that the RD remains unchanged in consecutive published lists for an inactive player although its value is less than 350. This is a result of rounding alone, RD continuously rises (or drops) vs time as the above graph depicts.

Calculations and publication of the rating list – the model of the Glicko system

The above examples of rating vs time and rating deviation vs time depict a real time observation of the changes caused by the completion of a game. A real time model can only apply to an environment of an ideal web based chess server where results of bullet and blitz games may occur any moment on a 24 hour per day basis. As soon as a new result arrives the server instantly updates the rating of the involved players. In this way the rating of a player is as accurate as possible whenever the server is requested by a user to output the current rating of a player. It is also possible that a user requests a rating list, not just the rating of an individual. The server's reply produces a list which is accurate only for a single moment, the one that the server handled the rating list request. A moment later rankings may change provided that a new result involving any of the listed players arrives.

This situation is typical for such a chess server operation regardless of the rating system used for calculations, ELO or Glicko. It is not possible for FIDE rating lists to follow changes in real time. Sometimes, tournament results are submitted with delay by the Federations' rating officers and there is always a possibility that a new rating list has been published between completion of tournament and results submission. This significant detail may cause problems when the official FIDE rating list is superimposed to rating lists calculated with a different rating system. As a result comparisons are not always accurate.

Furthermore, adoption of the Glicko system for the publication of a rating list on monthly basis requires a trick of altering the date of game completion, as explained below. This is also true for the ELO system, however the variable of time does not affect any parameter used for ELO calculations and therefore its importance is omitted when the ELO system is explained. The model of the ELO system can be described like this: The current ELO list was published January 1st early in the morning. A new list is expected to be published February 1st early in the morning. Since all data of the current period (January) are processed simultaneously by the rating algorithm we can safely assume that all games occurred at the same day. This day could be any day of the current period, but for comparison to the Glicko system, let's assume that this day was the afternoon of January 1st. From January 2nd to January 31st there was no chess activity. Submitted tournaments are processed in the morning of February 1st and the new ELO list is published as soon as the calculations are completed.

The significant difference regarding the Glicko system concerns the value of RD used in rating calculations. Although rating remained the same since a player's last result, RD must be assigned its current value before rating calculation, because its value changes continuously. The fact that RD must be correctly updated before rating calculations is an important detail of the Glicko system which must be taken into account for non-real-time applications such as FIDE ratings, where a new rating list becomes available every month.

The statistical theory utilized by Glicko system was developed to apply in many different fields, not exclusively FIDE tournament results. Every theory applies to a set of events that follow certain principles. In most cases an exact match of theory and its application is not possible, but proximity offers satisfactory results. In our case the use of the Glicko system demands that chess results must be adapted in a way that the theoretical model can be effectively applied. As a consequence, the model of the Glicko system, which alters the chronological sequence of events, must be taken for granted. It cannot be changed or modified.

Suppose that a Glicko rating list was published by FIDE in the morning of January 1st of the current year, stating the rating and RD of each player. This list provides accurate measurements of rating and RD as of the morning of January 1st. These data together with tournament results of the current rating period will be used in order to produce the next rating list expected to be published in the morning of February 1st.

In order to use the previously published RDs for the new calculations, all games of the current rating period must be considered to have occurred on January 1st. More precisely: the last list was published in the morning of January 1st and all chess activity occurred in the afternoon of January 1st. This is the only way that calculations can use a correct value of RD for each player. The remaining period from January 2nd to January 31st should be considered empty of any chess activity. Since the ratings remain unchanged during this empty period of time, all rating calculations are valid for the list of February 1st.

Although the calculations provided new values for the RD of each player, these values are only valid for the night of January 1st. The list of February 1st must contain new values taking into account the continuous change of RD during the inactivity period from January 2nd to January 31st. Updating RDs is necessary also for players who did not play any games from January 1st to January 31st. The formula is provided by Prof. Glickman but possibly it is not mentioned in the proposed reference.

After updating all RDs with the current values of February 1st, the new rating list of February 1st is ready to be published.

As mentioned above for the ELO system model, the date that all tournaments occurred could be any day of the current period also for the Glicko system. In fact real time calculations are possible but the complications of the formulas involved are out of the scope of this writing. The choice of January 1st in the above example is based on the suggestion of Prof. Glickman himself and there is an apparent practical reason for this choice: It allows calculation algorithms to use players' ratings and RD as published in the list of January 1st without any alteration of RD. This choice makes calculations less complicated and proves to be an important decision for the implementation of the Glicko formulas.

The Glicko system explained

Following the Glicko formulas manually is neither easy nor practical. Members of the chess community with little experience in computer spreadsheets (software like Microsoft Excel or Openoffice.org Calc) will find it difficult to implement the Glicko formulas on their own. Depending on other programmers' work, they will most probably face the Glicko system as a 'black box'. Thanks to Prof. Elo's ingenuity the 'black box' of the ELO system was reduced to the tables converting rating difference to scoring probability and vice versa. The Glicko system does not provide such facilitation. Scoring probabilities are dynamically produced during execution of the algorithm.

For the sake of simplification I am not going to repeat the exact formulas (which can be found in the first reference pdf on top of this document). I'd rather try to emphasize any resemblance with the ELO formula. So, the new rating of a player is calculated with the following equation:

$$R_{new} = R_{old} + K \cdot \sum g \cdot (score - E)$$

Does it look familiar? Of course it does. If you omit the factor 'g' from every summand of the summation operator and replace 'E' with P_D then the above formula looks exactly as the ELO rating formula. No one could have expressed this similarity more vividly than GM Bartlomiej Macieja: *"Arpad Elo's system can be compared to Newton's mechanics and Mark Glickman's system to Einstein's theory of relativity. That is, the ELO system is a subset of Glicko system."* Which conditions must be met so that ELO and Glicko calculate the same results? Well, the examples offered later try to provide an answer. For now, more information about the formula is necessary.

In order to calculate a player's new rating we need to know his/her current (old) rating and current RD from the last published list. For every game he/she played we need to know the opponent's current rating and RD as well as the score of each game. 'score' has the value of 1 if the player won the game, 0.5 if the outcome was a draw or 0 if the player lost the game.

'E' is the expected fractional score of a game and is a direct replacement of the score probability 'P_D' of the ELO formula. The value of 'E' is not taken from any table but it is the output of a function the arguments of which are the current ratings of both the player and his/her opponent as well as the opponent's RD. 'E' is always greater than 0, and less than 1. The limit values cannot be reached. 'E' depends only on the game in regard, not other games of the player.

'g' depends only on the opponent's RD for a certain game. It is the output of a function, the argument of which is the opponent's RD. If the opponent is an unrated player (so his/her RD equals 350) then 'g' has approximately the value of 0.669. A lower RD causes the value of 'g' to increase. 'g' is always lower than 1. It cannot reach the value of 1. It becomes clear that 'g' acts as a "weight" causing some games to contribute to the rating calculation more than other games. The Glicko system considers games against frequent players to be more important than games against players who compete rarely. Although the letter 'g' comes from the theory of probability, we may connect it with the word "gravity" because of its weighting function.

I use the letter 'K' in my presentation of the Glicko formula although this letter does not exist in Prof. Glickman's paper. Its function (not only its position in the formula) resembles the development coefficient used in ELO system, but it is not a constant. Its value is the output of a function, the arguments of which are the player's current rating and RD as well as every opponent's rating and RD. From all these arguments the one that affects its value more is the player's current RD. The coefficient factor of the ELO formula is determined by rating history and game history. Instead, the Glicko system's equivalent varies merely because of the player's rating deviation.

A frequent player has a lower 'K' compared to the 'K' of an unrated player. Its value may go as high as 485 for a rating period of an unrated player who has one game only against another unrated player. There are no "normal" values for 'K'. Its numeric value cannot be compared with ELO system's development coefficient. However, I presume that a "fine tuning" of the Glicko system will provide values varying from 5 to 20 for frequent players. This requirement of the Glicko system's implementation is addressed after some initial examples of calculations.

Test 1: ELO's score probability vs Glicko's expected score function

A characteristic of the score probability table of the ELO system is the fact that score probability varies according to the rating difference of the two opponents, not their exact ratings. The first test is about to determine that the same principle applies also for the Glicko system. Suppose that the two players have a rating difference of 70 points. This means that for the higher rated player P_D = 0.60 (ELO system).

R_{old}	R_{opp}	RD_{opp}	E
2400	2330	50	0.598
2200	2130	50	0.598
1900	1830	50	0.598
1600	1530	50	0.598
2400	2330	100	0.595
2200	2130	100	0.595
1900	1830	100	0.595
1600	1530	100	0.595
2400	2330	200	0.584
2200	2130	200	0.584
1900	1830	200	0.584
1600	1530	200	0.584

As you can see the output of Glicko's expected score function remains the same for two opponents with same rating difference regardless of the ratings. However, as said above, the 'E' value also depends on the opponent's RD. When this RD is different, then the value of 'E' also changes. Observe that the player, of which the rating is being determined, in our case the higher rated player, is expected to score worse when his/her opponent's rating is considered unreliable.

Notice that from the point of view of the lower rated player the results are supplementary as expected to be. According to this table, the lower rated player has somewhat better chances when his/her opponent's rating is considered unreliable.

RD_{opp}	E
50	0.402
100	0.405
200	0.416

Test 2: When the higher rated player outplays his/her opponent

The next series of tests try to determine the value of the development coefficient 'K' that the ELO system should use so that the rating change determined by the ELO system is exactly the same as the one calculated by the Glicko system. This hypothetical parameter can be found under the column K_{EQUIV}.

For this test, a 10 game match between two opponents is being rated, not just one game. This choice is consistent with Prof. Glickman's suggestion that the Glicko formula works better when an average of 5 to 10 games is taken into account for the rating period. It also allows testing of various result combinations that sum up to the same total score. However, this total score cannot match the score probability of the ELO system. In this case rating change would equal 0. Glicko rating change could also be 0 after rounding. This means that it would not be possible to determine K_{EQUIV} after setting the desired rating change in the ELO formula because the resulting equation includes an impossible division by 0.

A rating difference of 70 points was also chosen. The higher rated player's rating change is being examined.

R_{old}	RD_{old}	R_{opp}	RD_{opp}	score	E	P_D	ΔR_{ELO}	ΔR_{Glicko}	K_{EQUIV}
2800	50	2730	50	0.8	0.598	0.60	20	24.018	12.009
2800	50	2730	100	0.8	0.595	0.60	20	23.819	11.910
2800	50	2730	200	0.8	0.584	0.60	20	22.927	11.463
2800	100	2730	50	0.8	0.598	0.60	20	64.567	32.284
2800	200	2730	50	0.8	0.598	0.60	20	111.724	55.862

As you can see, although the opponents' RD varies (lines 1, 2 and 3) the Glicko rating change decreases only 1 (rounded) point. However if the player being rated is seldom competing then his/her rating gain increases compared to a frequent player, provided that he/she is able to outplay his/her 70 points lower standing opponent.

Although not pictured above, the Glicko rating system would provide the same results for a player of 2200 against an opponent of 2130 (and any rating combination with 70 points difference). However in this case ΔR_{ELO} would equal 30 because of the different coefficient factor.

Although not pictured above the Glicko system would provide the same results for any combination of scores that equals 80%. (+8 =0 –2) (+7 =2 –1) or (+6 =4 –0)

Let's examine the above table from the loser's point of view.

R_{old}	RD_{old}	R_{opp}	RD_{opp}	score	E	P_D	ΔR_{ELO}	ΔR_{Glicko}	K_{EQUIV}
2730	50	2800	50	0.2	0.402	0.40	-20	-24.018	12.009
2730	50	2800	100	0.2	0.405	0.40	-20	-23.819	11.910
2730	50	2800	200	0.2	0.416	0.40	-20	-22.927	11.463
2730	100	2800	50	0.2	0.402	0.40	-20	-64.567	32.284
2730	200	2800	50	0.2	0.402	0.40	-20	-111.724	55.862

Remember that rating deviation accompanies each player, so the colored lines will help you to examine the opponents in pairs. The pair of the first line (in both tables) defines a pair with equal rating deviation. In both systems the winner gains as many points as his/her opponent loses. However, when RDs are not the same then the players' gains and losses are irrelevant to each other. Take for example the red colored pair. The winner, whose rating is considered somehow unreliable, has gained 65 Glicko points from the match, while the loser lost only 24.

Notice once again the value of K_{EQUIV}. The high values of this hypothetical parameter indicate that some infrequent players when rated with the Glicko system may experience very steep rating changes (compared to the ELO system) if their performance is exceptionally good or bad.

Finally it must be noted that Glicko's "K' function has a different value (it is higher) than K_{EQUIV}. This happens because each game's contribution to the rating change has a weight determined by the 'g' function.

Test 3: Behavior of RD based on competition alone

The Glicko system formulas also calculate a new RD after the completion of data processing. Each rated game contributes to this calculation (the value of RD decreases because of game completion) but the value published in the next rating is increased in the end because of the inactivity period for the rest of the month. (This is going to be explained later.) The formula that calculates the change of rating deviation is not mentioned in this document. You can find it in Prof. Glickman's paper.

RD change because of competition is determined by the rating difference of the player and his/her opponents, and the value of RD of the player and his/her opponents. Each game contributes a unique change to the RD calculated. For this test a match between two players is to be rated for the current period, that is a number of games between opponents with the same rating difference and RD. The number of games determines a different average decrease of RD per game. This is demonstrated in the table below. The RD change per game can be found for 4 different lengths of the match. The score of each game does not affect RD calculated.

				Average decrease of RD per game for the given number of games rated in the current period			
R_{old}	RD_{old}	R_{opp}	RD_{opp}	1	2	10	20
2800	50	2730	50	-0.4786	-0.4719	-0.4246	-0.3784
2400	50	2330	50	-0.4786	-0.4719	-0.4246	-0.3784
2800	50	2730	100	-0.4479	-0.4419	-0.3998	-0.3584
2800	50	2730	200	-0.3548	-0.3511	-0.3241	-0.2962
2800	100	2730	50	-3.6719	-3.4835	-2.4982	-1.8712
2800	200	2730	50	-25.3099	-21.4695	-10.1319	-6.2772
2800	50	2600	50	-0.3674	-0.3634	-0.3346	-0.3050
2730	50	2800	50	-0.4786	-20.4719	-0.4246	-0.3784
2600	50	2800	50	-0.3674	-0.3634	-0.3346	-0.3050

From the above table you can see that there is no linear connection between the number of games and average per game drop of RD. You can also see that the decrease is the same for both players rated when their RD is the same. (Compare lines 1 and 8, then 7 and 9).

As rating difference increases then the RD change decreases in value. (Compare lines 1 and 7) This means that games against players with high rating difference are less important to the player's rating reliability.

Finally notice that a player who rarely competes can see his/her RD decrease in value with a fastest rate. The most important factor for rating reliability is indeed frequency of competition. Taking into account that the published RD is higher than the one calculated during game procession, the Glicko system seems to categorize players into RD intervals according to the frequency they play chess. This is not so clear at the moment. A closer examination of the increase of rating deviation because of time lapse will clarify things.

The global constant ‘c’ – how rating deviation is affected

A stage of preparation is mandatory before producing any Glicko ratings. It resembles to the tuning of a string because rating administrators must define the value of a single constant ‘c’ which is responsible for the steepness of the parabolic curve of rating deviation. Once defined, it cannot be changed. It is the same for all chess players since it is a characteristic of the system. A decision to change the value of ‘c’ after many periods of using the Glicko rating system would cause temporary instability of the rating system. Some players could experience unexpected changes to their rating.

This is a hard decision to take. Prof. Glickman suggests that a suitable value for ‘c’ can be determined by data analysis, a computing-intensive process. He also provides a practical method of setting the value of the global constant. This method explains adequately the essence of the global constant which is related to the desired number of periods of inactivity before the rating of an inactive player becomes as unreliable as the rating of an unrated player. It is the result of a simple equation:

$$c = \sqrt{\frac{RD_{UNR}^2 - RD_{NOM}^2}{t}}$$

The RD_{UNR} equals 350 following Prof. Glickman’s suggestion. ‘t’ is the number of periods before rating reaches the number of RD_{UNR}. RD_{NOM} is a value of rating deviation that is considered typical for the average chess player.

As soon as the global constant ‘c’ is set, the system is ready to calculate the exact value of rating deviation of the players which is going to be published in the next list. Prof. Glickman’s paper provides an equation for this calculation. The formula requires 2 extra variables which must be stored by the rating server (periods of inactivity and the RD from the last list that processed game results) for each player who shows no results for a certain period. Prof. Glickman provided a solution to this problem with the introduction of a formula that requires no extra data. In the following formula RD_{OLD} can either be the one published in the last list, when the player being rated is inactive for the current period, or the one calculated after processing games of the current period.

$$RD_{NEW} = \min\left(\sqrt{RD_{OLD}^2 + c^2}\,, \qquad 350\right)$$

It looks as if RD lies on the hypotenuse of a right angled triangle of which one cathetus is the last known value of rating deviation and the other one is the global constant ‘c’.

The following table provides the values of RD_{NEW} for different values of ‘c’ and for a different number of periods of inactivity. The initial value of RD_{OLD} is set to 50. So, 50 is either the value of RD calculated by the rating algorithm for the current period or it is the value of rating deviation taken from the last list if the player was inactive during the period. The following columns suggest that the player remained inactive for the number of periods indicated.

	Number of periods of inactivity								
	1	2	3	6	12	24	36	48	72
c=63.2	81	103	121	164	225	314	350	350	350
c=31.6	59	67	74	92	120	162	198	223	271
c=22.3	55	59	63	74	92	119	143	167	191
c=15.8	52	54	56	62	74	91	103	115	139
c=7.95	51	52	53	56	62	63	63	63	63

The first line has such a high value of ‘c’ so that RD reaches 350 after 30 periods. When c=31.6 the value 350 is reached after 120 periods. When c=22.3 the value 350 is reached after 240 periods. When c=15.8 the limit superior should be reached after 480 rating periods. With c=7.95 the number of periods becomes 1898. Notice the line for c=7.95. It seems that RD will never increase more than 63. Of course this is misleading! A system that stores RD values as integers can no longer increment RD value, because of rounding. This detail should point out that implementation of mathematical formulas may accidentally introduce errors (such as rounding to the nearest integer) that might become chaotic when theoretical results cannot be confirmed.

The above formula allows the production of a table (for a given ‘c’) that shows RD change after one period of inactivity. For a consecutive inactive period we can find RD_{NEW} from the corresponding column of the table. Part of this table may look like this for c=31.6

RD_{OLD}	…	50	51	52	53	54	55	56	57	58	59
RD_{NEW}		59	60	61	62	63	63	64	65	66	67

RD_{OLD}	60	61	62	63	64	65	66	67	68	69	…
RD_{NEW}	68	69	70	70	71	72	73	74	75	76	

Tuning of the Glicko system

A great characteristic of the Glicko rating system is its ability to adapt to the special needs of the system it applies to. This means that a chess server where most of the games are blitz and bullet may have a different implementation of the Glicko formulas compared to a correspondence chess organization. Chess servers that provide “live ratings”, where a player can play many blitz and bullet games, may process many games per player per day, so a value of ‘c’ less than 8 might be more appropriate, while RD should increase on a daily basis. For a system like FIDE rating server, one period of inactivity should be defined as the amount of time between consecutive publications of the rating list. With average tournament duration of

10 days and delay time until tournament submission, a monthly list is probably the best response one could expect from FIDE.

Austin Lockwood, the system administrator of correspondence chess site schemingmind.com, is already using the Glicko system for the past 10 years. From his own perspective, the choice of the right period between two consecutive lists was the first problem he faced. He decided that a period of one month would maintain a dynamic system. A longer period would result in player frustration. Choice of ‘c’ was the second problem he had to solve. In his own words:

“Initially the value of ‘c’ was chosen based on the recommendations in Prof. Glickman's paper, assuming that a player’s RD would change from 50 to 350 following five years of inactivity (60 periods, or c = 44.72). Unfortunately this seemed to result in active players having high RDs and rather volatile ratings, in some cases new players were obtaining extreme ratings after only one or two games. In addition, it's counter-intuitive to assume that a correspondence player's rating would become uncertain after five years of inactivity. We didn't want ratings to become completely viscous, so we were cautious in decreasing ‘c’. Eventually we decided to assume that the period in which the RD of a player with RD=50 increased to 350 would be ten years (or 120 periods), we therefore currently use a value of 31.6 for c. To date, this seems optimal, as there is still fluidity in the ratings, but fewer anomalies.”

Jeff Sonas experimented with the actual set of data submitted to the FIDE rating server. He implemented the Glicko system and tested different values of ‘c’ always having in mind Prof. Glickman’s suggestions. Here are some of his comments taken from the last reference on top of this document:

“First of all, I decided to just use Mark's example value of 63.2 in my own implementation of Glicko. This would suggest that if a typical player stops playing for t=30 months (i.e. 2.5 years), their rating uncertainty should be comparable to that of an unrated player. I think this is way too short; even after 2.5 years I would still have more confidence in the idle player's old rating than in that of an unrated player.

I think that a value of t=120 would be more appropriate, maybe even a bit low, saying that it is only after a player has been idle for ten straight years (120 months) that we consider their rating to be equally uncertain to that of an unrated player. So, solving for c as per Mark's suggestion, we get a value of c exactly half of what his example yields, so a sensible value for me to have used would be c=31.6. However, in my code to implement Glicko, I forgot to square the value of c when updating rating deviation, so I had effectively used c=7.95 (the square root of 63.2). Following Mark's approach for the meaning of c, this would imply a player would have to be inactive for 160 years until their rating deviation would finally reach comparable uncertainty to that of an unrated player. Most likely this value of c is a bit too low!

So I have now tried all three of these values of c, in the Glicko system. c=63.2 would be the lazy choice for c (in which ratings would probably get too uncertain too fast), and c=31.6 is a more reasonable choice for c. When I tried these, I found that c=7.95 still did the best of the three, though c=31.6 did almost as well and certainly way better than c=63.2. Since the "mistaken" value of c=7.95 seems to perform best, that suggests that we really don't want ratings to degrade very rapidly. So I tried one more option, right in the middle *(geometrically), of the two best, namely c=15.8 (so c-squared = 250). This would imply that a player would have to be inactive for 40 straight years before their rating is as uncertain as an unrated player - surely this is still too low a value of c? But it worked the best of the four, in my own validation, and so that is my recommended value of c (c=15.8, the square root of 250). Possibly this just implies that 350 is not an appropriate choice for the RD of an unrated player, but that seems like enough optimization for me."*

I have no further comments upon the choice of the global constant 'c'. Instead of proposing a value myself, I decided an approach from a different angle. By using test data instead of real tournament results I tried to determine the consequences of such a choice. Manufactured data containing convenient results revealed valuable information for the Glicko system. All my conclusions are related to the choice of 'c'. So, I ran tests for c=15.8, c=31.6 and also for their geometrical middle c=22.3.

The agenda of the tournament player

Now that we have examined the way that RD decreases because of game activity and increases because of inactivity, it is time to comment on the consequences of the choice of the constant 'c'. After many periods of processing test game results (only updating player's rating and RD) rating deviation obtained a stable value which depends only on the number of games per period and the global constant 'c'.

The idea of RD=50 for a typical player comes from Prof. Glickman's paper. One would expect that an infinite repetition of the same results would provide such an accurate rating measure so that rating deviation reaches a minimum value of 1 or 2. This is totally wrong! The typical RD value is not a property of the system. It is rather a characteristic property of the player, a result of his/her playing habits. The following table tries to determine this optimum value of RD depending on a standard number of games per period and the global constant 'c'. The table is based on a test designed to determine an unrated player's path to reliability. It is explained later.

	Stable value of RD reached		
Number of games per period	c=15.8	c=22.3	c=31.6
1 game	81	94	111
5 games	54	63	75
10 games	45	54	65
15 games	41	49	59

It is possible for a player who plays a standard number of games in each period to maintain his/her rating deviation at a stable level. It is possible that periods of inactivity are inserted between active periods because of the player's agenda. Still, rating deviation can be located within a limited interval of values provided that the player's chess habits do not change year after year. Maintaining rating deviation is not a goal, of course. It is a result of programming participation in chess tournaments.

One step further ahead (yet to a dark destination), I tried to determine the value of K_{EQUIV}. May I remind you that K_{EQUIV} is a hypothetical development coefficient that would allow ELO system to provide the same results as Glicko. The K factor of the ELO system was more likely a property of the system based on the history of the player's ratings. The Glicko system's K function provides a K_{EQUIV} which is merely determined by the player's playing habits. It's the player who decides the ratio that his/her results affect his/her rating change. The tests showed that K_{EQUIV} has a very fluid value. As a result, the contents of the following table are indicative rather than accurate. However, even if an error of ±2 is considered, the table still demonstrates the ability of a player to control his/her fate.

	K_{EQUIV} for RD_{OLD} taken from the above table		
Number of games per period	c=15.8	c=22.3	c=31.6
1 game	33	46	60
5 games	13	19	26
10 games	9	14	18
15 games	8	10	12

It is often said that some players compete rarely in order to protect their rating. Since their K_{EQUIV} is higher because of their playing habits, an unexpected bad performance might result in a big loss of points. I hope that this observation might encourage some players to play chess regularly, not make them disappear from the playing halls.

FIDE rating lists may provide enough data for the habits of the chess players, as the following table shows. Considering rating zones interval of 200 points, it is evident that players do not have the same opportunities to play FIDE rated chess. Professionals with rating higher than 2400 play more frequently than lower rated players. Although the table is too fragmentary we can also conclude that chess activity is not equally distributed throughout a year. Also notice that the rating periods pictured are two months long. Up to this date, the monthly list has not become available yet.

	Players who concluded games in each rating period			Average number of games per player submitted to FIDE		
rating	Jan11	Sep10	May10	Jan11	Sep10	May10
2601-	122	153	175	10,76	16,03	16,28
2401-2600	954	1222	1391	11,29	16,88	13,34
2201-2400	3032	3469	5371	8,76	12,57	9,66
2001-2200	7116	7264	12172	7,12	9,89	7,62
1801-2000	7832	8311	11156	6,87	8,98	7,07
1601-1800	5572	5668	5820	6,5	8,59	6,55
1401-1600	2089	2134	1749	6,6	8,41	6,46
1200-1400	337	301	137	7,22	9,43	7,33

From the above table we cannot readily conclude that chess players have an average of 5 games per month. A more detailed analysis is required in order to determine the average number of submitted games per period combined with periods of rest throughout a year. When these data become available we will be able to reach a better approximation of typical rating deviation values for each zone.

The case of the unrated player

FIDE's handbook chapter 8.0 "The working of the FIDE rating system" spends more than half of its length to describe the unrated player's entrance to the system. The downloaded text file of FIDE rating list including all players (rated or unrated) contains 285044 players (as of May 24th, 2011), 54% of which are unrated players. Almost 155000 players have participated in a rated tournament (and they received an ID by the system) but their rating is not available yet because they don't match the criteria to get an initial rating.

Prof. Glickman suggests that an unrated player receives 1500 Glicko points and RD=350 as soon as he/she plays for the first time in a rated tournament. That's great news because there are no unrated players anymore! But how long does it take before this player is considered an established player? Of course he/she must play chess on a regular basis. Suppose that an unrated player is able to play at a consistent level and that his/her results continuously confirm his/her ability. The following test tries to determine the Glicko system's reaction to this steady performance.

An opponent of constant rating throughout the rating periods is chosen, against whom the unrated player can always achieve a 50% score. This implies that after a number of periods the unrated player can reach the rating of his/her opponent. The test was repeated for three different targets: 1700, 1900 and 2100 Glicko points. Also three different values of 'c' were chosen and three different number of games per period.

	c=22.3, 10 games played per period					
	target=1700		target=1900		target=2100	
Periods	rating	RD	Rating	RD	rating	RD
0	1500	350	1500	350	1500	350
1	1715	124	2144	169	**2942**	**235**
2	1707	86	1949	110	1611	220
3	1704	72	1925	82	2256	163
4	1703	65	1916	70	2143	101
5	1702	61	1912	63	2124	79
6	1702	58	1909	59	2116	68
7	1702	56	1907	57	2112	62
8	1702	55	1906	55	2109	59
9	1702	54	1905	54	2107	57
10	1702	54	1904	54	2106	55
11	1702	54	1903	54	2105	54
12	1702	54	1902	54	2104	54
13	1702	54	1902	54	2103	54
14	1702	54	1902	54	2102	54

Knowing the unrated player's level, I consider the Glicko's rating reliable when it has a deviation less than 1% of the target value. (The term deviation is used literally here and has nothing to do with the RD.) For the above targets, reliability arrives after 2, 4 and 6 periods. The closer the target level lies to the initial rating of 1500, the fewer periods the system needs to reach the target. The green area marks the periods during which the rating continues to improve towards the target until rounding to the nearest integer does not allow further improvement. RD value becomes stable in the end as it was mentioned before.

The cells colored red under the 2100 points target provide a good idea of Glicko system's respond, an oscillation around the target value, the amplitude of which decreases period after period. However it is possible that an unexpectedly high value is reached (in this case r=2942, RD=235) because the player performs much better than 1500. Although the values are correct and must be used unedited for the next calculation, some action is required in order to prevent an erroneous rating from publishing.

	c=15.8, target rating =1900					
	5 games		10 games		15 games	
Periods	rating	RD	Rating	RD	rating	RD
0	1500	350	1500	350	1500	350
1	2023	215	2144	169	2197	144
2	1940	133	1949	109	1938	96
3	1923	103	1925	80	1918	68
4	1916	88	1917	67	1912	57
5	1912	79	1913	60	1909	51
6	1910	72	1910	55	1907	47
7	1908	67	1908	52	1906	45
8	1907	64	1907	50	1905	43
9	1906	61	1906	48	1904	42
10	1905	59	1905	47	1903	41
11	1904	58	1904	46	1902	41
12	1904	57	1903	45	1902	41
13	1904	56	1903	45	1902	41
14	1904	55	1903	45	1902	41
15	1904	54	1903	45	1902	41

The second table compares a different data set. This time the target is fixed but the number of games per period changes. Obviously more games per period means that rating becomes reliable sooner. Notice that RD decreases rapidly towards its optimum value. It is quite impressive that for 5 and 10 games per period a reasonable value for rating parameters is reached after 4 periods although the number of total games is 20 and 40 respectively. It seems to me that a player who has the opportunity to play one tournament per period for 4 periods will be able to be considered an established player in a shorter period of time.

The last table appears to be exceptionally interesting. The target value is reached quite soon, only after the third period (where the initially unrated player has played only 15 games total) but RD continues to decrease until it reaches the typical value for 5 games per period for the given 'c'. Rating oscillation accidentally (?) reaches the target value too soon and after this point it approaches closer and closer to the target. Notice that all 3 different values of 'c' calculate almost the same rating. However, RD decreases with a different rate. The lowest 'c' value drops RD faster but reaches its final stable value with a longest delay.

This unusual behavior of the Glicko system which allows some players to obtain high ratings for one period could possibly whet the appetite for result manipulation because newcomers can metaphorically be “pinned” to their actual strength level without having to play many tournaments. This is of course a highly improbable but not impossible scenario.

	Target rating=2100, 5 games per period					
	c=15.8		c=22.3		c=31.6	
Periods	rating	RD	rating	RD	rating	RD
0	1500	350	1500	350	1500	350
1	**2497**	**276**	**2497**	**276**	**2497**	**277**
2	2075	193	2075	194	2074	195
3	2090	124	2090	125	2090	127
4	2094	99	2094	101	2094	104
5	2096	85	2096	88	2096	93
6	2097	77	2097	80	2097	86
7	2098	71	2098	75	2098	82
8	2098	67	2098	71	2098	79
9	2098	64	2098	69	2098	77
10	2098	61	2098	67	2098	76
11	2098	59	2098	66	2098	75
12	2098	58	2098	65	2098	75
13	2098	57	2098	64	2098	75
14	2098	56	2098	63	2098	75
15	2098	55	2098	63	2098	75
16	2098	54	2098	63	2098	75

I believe that the unrated player should be treated according to theory, but his/her rating should not be published (of course rating and RD are stored and used for next period calculation) from the first period with submitted data. Perhaps the second or third period should be preferred. Perhaps another criterion could be used, for example have his/her RD lower than 200 so that he/she is considered an established player. Number of games is not the best choice here since it is possible that a newcomer reaches a rating objectively too high. But this is FIDE’s problem, not Glicko’s.

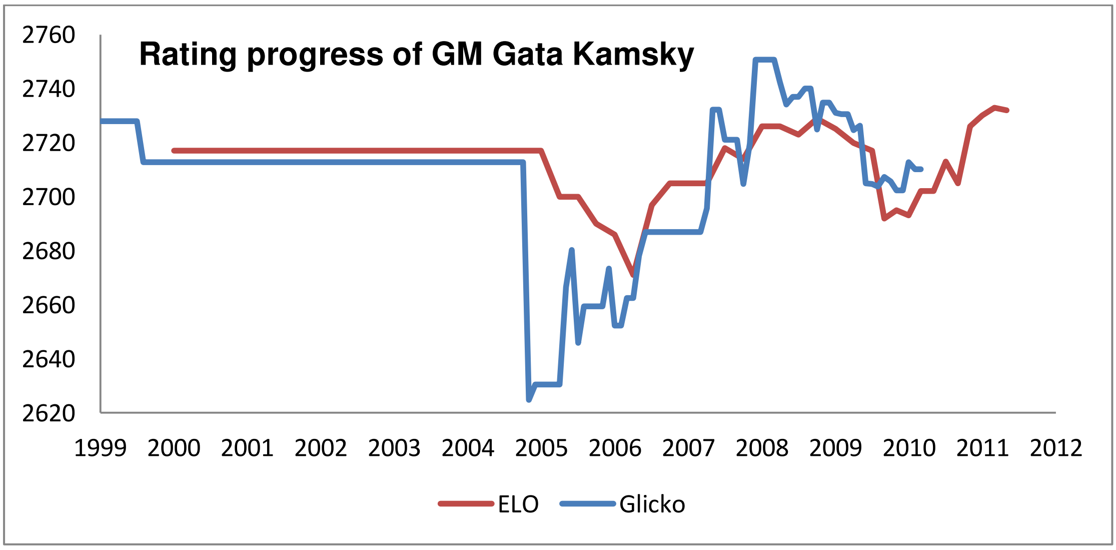
An unrated player requires a number of periods before his/her rating represents his/her playing strength with confidence. This is easy to understand. However an unrated player brings to mind a youngster who has not tested his/her chess ability on the board. It is difficult to compare an experienced but inactive player with an unrated one. Yet, the Glicko system handles both in a similar way.

The case of GM Gata Kamsky

GM Gata Kamsky is a showcase example of a strong player who returned to chess after many years of inactivity. His ELO rating chart shows a loss of 46 points which he was able to gain back after a period of 2 and a half years and 127 rated games. I decided to test his Glicko rating progress and compare it to the ELO rating calculated by FIDE. Not being able to perform an accurate Glicko measurement (because I should use Glicko ratings and RDs for all players he played against) I asked Jeff Sonas to help me.

An important technical problem does not allow a perfect match of the Glicko rating graph pictured together with the FIDE ELO graph. Jeff Sonas processes the data as soon as games are completed. Games are included in the list of the month they occur. As a result a tournament may be split in two consecutive lists. For example this happened with the 2005 US Chess championship, the first tournament that GM Kamsky participated in after his long absence, half of which was evaluated in the November 2004 list, while the other half in the December 2004 list. However, the tournament was submitted to FIDE on January 28th, 2005, so its results were processed in the April 2005 FIDE rating list.

Even so, the results of the comparison pictured below are quite impressive and indicative of the Glicko system's performance.



The ELO system reacts to the results with a "V" line. GM Kamsky loses ELO points because his performance is worse than his rating. When his performance improves his rating also improves. On the other hand, the combination of bad results (not consistent with rating) and a high RD lead to a big loss of Glicko points as soon as the player returns to activity. The vertical drop causes a loss of 88 points. Afterwards, his performance improves tournament after tournament and Glicko system reacts to his results with an oscillation that follows this improvement. I expect that Mr. Kamsky can confirm that after he returned to tournament competition he was playing better chess day after day. This is exactly what the above graph presents.

Performance matters

I avoided mentioning the word performance until the GM Kamsky rating progress example. Rating performance is a very useful measure for the players as it indicates how well they played, especially in a chaotic Swiss tournament when lost somewhere in the middle of the score table. Although performance rating has nothing to do with rating change, it plays an important role for title applications. FIDE handbook provides a formula for calculating performance rating but the definition of the term is missing. Two different definitions can be found in Wikipedia (under “ELO rating system” and “Glossary of chess” lemmas) but none of them is correct.

Performance rating of a player in a certain tournament is the hypothetical rating of the player before the tournament which would result to a rating change of 0 points after the same tournament results. The FIDE formula for performance rating calculation confirms this definition in most cases. However, when the player has unexpected results against opponents with a high rating difference, it is possible that the FIDE formula and the one derived from the above definition provide different results. In addition, there are more than one values of performance rating that confirm the above definition (because of rounding to the nearest integer) but only one of them provides the best proximity. For example, GM Vladimir Kramnik’s results from the recent Khanty Mansiysk Olympiad: His performance rating is calculated to 2794 using the FIDE formula, but all values from 2794 to 2800 also confirm the definition, with 2798 points being the best choice.

Prof. Glickman gives no formula for performance rating. This is the reason why a good definition of the term is required in order to produce the correct formula from Glicko system theory. If we set $R_{new} = R_{old}$ in the Glicko rating formula, we get a new equation: $K\Sigma = 0$. We need to determine the value of R_{old} (in our case this variable represents the performance rating) which is a variable of both factors of the product. However, only one of the factors should equal 0. Since ‘K’ is a function, an argument of which is the player’s RD, the condition K=0 can only produce a result depending on the player’s RD. So, we must solve the equation Σ=0 for R_{old}. The solution requires an algorithmic approach.GM Kramnik’s Glicko performance in Khanty Mansiysk Olympiad is calculated to 2797 points.

Unfortunately, the use of a programming method for the calculation of Glicko performance rating is inevitable. An effort of solving the above equation manually leads to a polynomial equation, the degree of which equals the number of games played in the tournament, a very difficult task if not impossible. The same applies to the calculation of the ELO performance rating according to the definition provided. The formula described in FIDE handbook (1.48 – 1.48a of the International Title Regulations) is not free of error, but it results to the best possible approximation and allows a quick manual calculation. Of course, it is possible to use the same formula for Glicko performance rating. The formula can be further improved by slightly changing the conversion table so that the proposed rating difference matches more precisely the percentage score observed in tournaments.

Glicko as the successor of the ELO rating system

A possible decision of FIDE to abandon the ELO rating system for a new one (not necessarily the Glicko system) requires a very careful examination of the available systems by the Qualification Commission. All systems being examined should provide more accurate rating measurements than the ELO system. So, the first task of the QC is to **decide a method for testing and comparing rating systems**. Jeff Sonas has already created one and used it for evaluating the systems submitted to the Kaggle competition. According to Jeff, the Glicko system is 6.6% better than the ELO system, but the Stephenson formula of the Glicko system (I call it like this because Dr. Alec Stephenson of Swinburne University, Australia, has provided a well-tuned and carefully prepared variation of the Glicko system which takes into account an extra parameter: the color of the pieces) is 10.4% more accurate than the ELO system. If 6.6% doesn't seem impressive, then 10.4% certainly does! The Sonas method for testing the rating systems is not compulsory; it is the only one available for the time being. Any method approved by the QC can be used for this purpose.

The second task of the QC is to **compile a catalog of all the problems associated with the ELO rating system**. A public dialogue might also be helpful, as it is possible to highlight issues that escape our attention. The commission should examine if the listed problems continue to exist when the new system is implemented. Since a complete database of tournament results from the past 10 years is available, **a retrospective ratings calculation is required** using every potential successor of the ELO system. These lists may provide valuable information about the working of the new rating system. Comparison of individual rating progress of players with different playing skill will become available. The publication of these lists can help players worldwide to evaluate the new rating system by comparing their individual progress to their ELO rating progress.

Retrospective calculations can also be useful for the evaluation of the Title Regulations too. **A sufficient number of norm certificates of pending and old titles must be reexamined** using opponents' ratings from the retrospective lists in order to determine possible slight adjustments to the ratings required so that the playing strengths for achieving any title are preserved after the adoption of the new rating system.

Apart from the above general tasks, the use of the Glicko system by FIDE requires some extra work by the QC. Let's summarize all Glicko specific decisions that must be taken:

- The choice of the global constant 'c'.
- The precision of the rating deviation value (integer or floating point) and the method for calculating the rating deviation of the inactive player.
  (The significance of this decision is not analyzed above since it was realized in the final preparation stage of this paper.)
- The frequency of publication of the Glicko rating list.
  (I consider that a monthly list is the optimum choice.)
- Initial ratings and rating deviation used for the calculation of the first Glicko rating list.
  (I propose that the ratings from the last ELO list should be used, but rating deviation should come from the retrospective rating calculations.)
- The conditions met by a new player (the unrated player of the ELO system) so that his name is included in the lists (provisional rating).
- The conditions met by an inactive (or an incidental) player so that his name is omitted from the lists.
- The formula (or method) for performance rating calculation.

Finally there is another problem (perhaps not the last one) which I avoided to mention because I could not find a solution. It will not be possible to process late submitted tournaments using period ratings. The ELO rating system is so intelligently simple that late tournaments can be incorporated without significant error. (It must be mentioned that the procedure followed today is not free of errors.) On the contrary the Glicko system requires that all the player's results from the rating period must be used in order to calculate new ratings and rating deviations. Glicko rating calculations cannot be done incrementally.

Although calculation of ‘live ratings’ is still possible (sites providing live FIDE ratings will not lose their audience) a late tournament processed in the correct period might cause a chaotic recalculation of ratings since all the players involved in this tournament and their opponents’ from future periods need to be recalculated.

We cannot disregard late submitted tournaments. Processing them in the current rating period introduces significant error in the calculations, but the most important problem is going to be the complaints from the players. Is it possible that all tournaments are processed in the correct (the current) rating period? The answer is ‘yes’ if all the tournament results arrive on time. I only want to say that the Glicko rating system, if used by FIDE, will be more closely connected with the informatics technology infrastructure that provides the rating service. It will be necessary to upgrade the tournament submission procedure. If so, then why not proceed to a carefully prepared radical reform of the FIDE rating service?

“Rating IS a commodity!” (FM Aviv Friedman)

Many thanks to

Prof. Mark Glickman for his suggestions and approval of my work

Jeff Sonas for his immediate response when I asked him to help me

Austin Lockwood, without whom I wouldn’t have found the stimulus to study the Glicko rating system.