Quality of play in chess and methods for measuring

Erik Varend

Tallinn, 2014

Abstract. In this study, using the computer, the subject of the research is the absolute strength of play of various chess playing entities (humans and computers). First of all, the actual accuracy of play will be determined which is measured via the mean difference between the move suggested by the engine and the move actually made. Thereafter, individually for each entity, factors that have an effect on the accuracy of play will be determined, and an estimated accuracy of play will be found based on those factors. It shows the accuracy if all factors were the same for all players. As a result, it was determined and proven that there is a relationship between rating and the quality of play. In addition, it was also proven that the further one goes back in time, the more the quality of play decreases. By comparing the accuracy of play in both humans' and engines' play it was determined to what extent CCRL and FIDE ratings correlate. The author also drew several miscellaneous conclusions based on the collected data.

1. Introduction

The primary aim of this study is to find a correlation between the strength of play (either FIDE and CCRL) and the accuracy of play. Also 7 most noteworthy performances in the history of chess are under comparison, and how the strength of chessplayers has changed over time. Besides, the final section of the paper contains various other conclusions that can be drawn from the data collected.

There are 2 different ways to estimate and compare performances:

1 by measuring absolute strength;
2 by measuring relative strength.

Absolute strength can be defined by how far away a performance is compared to the perfect performance, i. e. the distance between the actual performance and the best performance possible. The closer a performance stands to the absolute perfection, the better it is. In case of relative strength a performance is compared to results of other performers, and the actual strength has no importance at all. In circumstances where ascertaining the absolute strength has not been easily feasible, there has normally been no choice but to use relative strength measurement as a yardstick. The latter is prevalent in case of one-on-one sports such as snooker, tennis and chess where ELO rating is used to compare the strength of players. Comparisons of players and performances from different epochs is only possible using absolute strength. It is, for example, impossible to say who was stronger, Lasker or Spassky if using their chessmetrics ratings.

Consequently, we need to find an indicator of absolute strength in chess. There is a variety of ways to do this which can be split into two primary types:

1 tablebases
2 various computer-based estimations

Certainly, most preferred would be tablebases, because they give perfect solutions for each position. In that cas the accuracy of play would be measured in the mean number of transitions per move. A transition is a change in the state of the game – won, drawn, or lost – assuming perfect play from both sides. A change from a drawn position to a lost one, or from a drawn one to a lost one equals to 1 transition. If a won position becomes a lost position, it is 2 transitions. The fewer transitions per move, the higher the quality of play. Four piece tablebases were completed by the end of 80-s. 5-piece TBs were compiled in early nineties, those with 6 pieces in 2005, and now we have 7-piece tablebases. This

implies that it's quite hopeless to see the complete 32-piece tablebases in near future. That's the reason why chess engines are necessary. There are many ways to describe the absolute accuracy with the help of the chess engines:

1 The average difference between the best move suggested by the engine and the move actually made
  1.1 difference expressed in centipawns
  1.2 difference expressed in percentages
2 The average change in evaluation after the move made by the player
3 The percentage of moves that coincide with those suggested by the engine
4 The percentage of moves where the error exceeds a predetermined threshold

The version 1.1 can be called the classical method, since it was used by Slovenian researchers I. Bratko and M. Guid in their groundbreaking study.[1] The magnitude of an error is essentially the centipawn gap between the evaluation of a move suggested by the engine and a move actually made. Smaller differences indicate more accurate play.

Another promising possibility is to use percentages instead of centipawns, i. e. similar to Monte-Carlo method. The percentage indicates white's scores against black after a move. To find the score, a computer is set to run a certain number of games against oneself. Better scores would represent moves that are more preferable. The downside is the fact that it takes a lot of time to get a statistically valid number of moves, especially taking into account the need for ensuring that the engine has enough time per move. Otherwise its useflness in more complicated positions becomes questionable due to the horizon effect. Its advantage primarily lies in theoretically drawn endgame positions where evaluation-based estimations are known to be unreliable. When a position is 100% drawn, Monte-Carlo method always shows 50% score, while evaluation may, in some instances, be over 3.00. Therefore, there is always a risk that an evaluation-based estimation may detect false 'blunders', assigning large evaluation differences to moves which are actually drawn.

A Portuguese scientist D. R. Ferreira has worked out an interesting alternative solution, where what matters is not the gap to the best move at the same position, but between evaluations of best moves before and after a move has been made by a player.[2] Like the classical method, Ferreira's method can be used with percentages.

These tables below display differences in the classical and Ferreira methods in the cases of centipawns and percentages.

move	evaluation	gap	Move after Ne5	evaluation	change
13.Ne5	0.34		13...h6	0.01	-0.33
13.Rc1	0.31		13...Bc4	0.04	
13.e4	0.05	0.29	13...dxc4	1.03	
13.Nc2	-0.12		13...f5	1.34	
13.a3	-0.14		13...Bxe5	1.80	

move	percentage	gap	Move after Rc1	percentage	change
13.Rc1	55%		13...h6	52%	-3%
13.Ne5	50%		13...Bc4	55%	
13.e4	50%	5%	13...dxc4	93%	
13.Nc2	40%		13...Bxe5	97%	
13.a3	39%		13...f5	99%	

The ways 3 and 4 are clearly inferior.
The fact that a move made by a player coincides with that of made by computer may in most cases indeed indicate a good move. Whereas non-coincidence between computer's and player's moves does not say about the quality of a move. How important it is to play the best move according to the engine, depends on a particular position; in matters more in the case of forced positions. In positions where there are a lot of more-or-less equal moves it is not possible to determine the quality of a move by using that method.

The disadvantage of the threshold method is that it treats all moves below a certain threshold as being equal to 0.00. If, for example, the value of a threshold is chosen as 0.25, then, in case of a player making all of its moves in 0.20-0.24 range, its play would be considered as perfect. If a player makes 4 moves where the error is 0.20, then it makes as big

1 http://www.ailab.si/matej/doc/Computer_Analysis_of_World_Chess_Champions.pdf
2 http://web.ist.utl.pt/diogo.ferreira/papers/ferreira12strength.pdf

impact as one error of 0.80. The principal problem of the both methods lies in the fact that they are too coarse and do not describe the position of moves on the quality spectrum.

However, deterimining the absolute accuracy of play alone is not sufficient. A performance never happens in a vacuum, isolated from all factors acting upon it. The level of a performance can only be manifested by the co-influence of the two factors:

1 potential
2 conditions

Potential is the ability of a player to exhibit an as high standard of performance as possible. It depends on a variety of characteristics that differ for each sports. For instance, physical sports require good physique, stamina and technical skills. In mental spords, such as chess and go, the required characteristics would include short-term memory, calculation speed, intuition etc.

Conditions refers to a set of factors upon which the accuracy of play depends:

1 difficulty of positions
2 thinking time
3 practical play
4 psychology
5 conditions in the venue
6 health
7 level of fatigue

The first three ones are the most important.
In some positions it is easier to find a good move, whereas in other positions it is more difficult. That's what the term 'difficulty of positions' refers to. There are many ways a position can be difficult, it cannot be described by a single factor alone. It consists of many aspects, for example, it may be chaotic and complicated, or there are relatively few good moves in a position, or good moves appear illogical at first sight etc. Also, difficulty is individual and varies among different players: what for one player is difficult, may be easier for another. Computers generally are able to find illogical moves with greater certaincy than humans.

Thinking time is just a time control games are played under. Over time rate of play has gradually gotten increasingly shorter.

The notion 'practical play' refers to the phenomenon where a player intentionally sacrifices the accuracy of play to make matters more difficult for the opponent. The goal is to create such a situation where he would have to make comparably more effort to maintain the same level of accuracy. There are 3 kinds of situations that can be perused in practical play.

1 difficulty of positions;
2 suitability of the type of positions;
3 thinking time.

Suitability of the type of positions indicates how much a certain type of positions suits a player and his nature; whether he is familiar with such type of positions, whether a given position needs more of calculating, knowledge, intuiton etc. In the start position and usually at the beginning phase of the game all the three factors are even for either player. The aim is to introduce imbalances into the game situation, in favour of the first player itself, so as to the opponent has more difficult positions which also are less suitable for him. If the opponent is in time trouble, then moving faster so that he has less pondering time.

Psychology plays an important role in chess. A chess player must have willingness to endure competitive stress. It is important that he has ability to remain calm in critical moments. Sometimes it happens that a chess player allows himself to be disturbed by psychological factors, such as problems in private life, concerns over homeland or relatives and friends, that can affect concentrating on the game, and hinder going all out. The third type of psychological factors is directly connected to chess; whether incompatibility with the style of an opponent, fear, or a feeling of uneasiness with him. Probably the most famous example is Shirov's lifetime score against Kasparov – 7:22 with no wins, which is far more than one could expect from their ratings. Also, one may have gotten used to the style of a particular chess player to the extent that unexpected sudden changes in his play may confuse. Among these are cases where a player, who usually has preferred correct and objective play, suddenly sacrifices material. Believe him or not?
Consequently, psychological factors can be broken in three main types:

1 factors arising from player's characteristics;

2 external factors;
3 chess-related factors;

The rest four factors are less important, but they are not to be ignored.
Naturally, conditions in the playing venue must be as good as possible. A chess player has to be allowed to concentrate fully on his game. Every disturbance may affect negatively on play. In today's chess, conditions in the venue of top tournaments may be regarded as luxurious. But it wasn't common in the past. On older photographs, it can be seen how lighting at the venue leaves much to be desired, there were hardly enough space on tables for scoresheets, spectators were crowding near players etc. Such things would be unheard of nowadays.

Good health is also a prerequisite for a decent performance. Feeling sick as well as being under the influence of substances undoubtedly drag down the quality of play.

It's the same story with stamina. Older players are more susceptible to it. Tiredness primarily occurs at the final phase of game; mistakes stemming from exhaustion can be found more often in the endgame than rest of the game phases.

It's evidential that only the first three (difficulty of positions, time control and partially practical play) can be a subject to being measurable. This time there are no known ways to measure tiredness, psychological stress, conditions in the venue etc, and their effect on players.

The relationship between the accuracy of play, conditions and potential is illustrated by the formulas below:

$$accuracy = \frac{potential}{conditions} \qquad conditions = \frac{potential}{accuracy} \qquad potential = conditions * accuracy$$

Note the similarity with the Ohm's law:

$$current = \frac{voltage}{resistance} \qquad resistance = \frac{voltage}{current} \qquad voltage = resistance * current$$

On the ground of this analogy, it can be asserted that attempts to determine the strength of play of players without taking into account conditions moves are made in, is the same as trying to find voltage without knowing resistance and current.

In conclusion, it is clear that before one sets about performing analysis, one has to pick one of the methods for describing the accuracy of play and find methods to measure the effect of the difficulty of positions, time controls and practical play on it. Also, it is essential to reckon with the magnitude of evaluations that serve as the basis of describing the accuracy of play. The next part touches that subject.

2. Methods

The research focused on three problems:

- investigate and map the increase of the absolute level of play throughout the history of chess
- find out the correlation between FIDE and CCRL (computers) rating sytems and the level of play
- compare the 7 most remarkable performances in the history of chess
  - Emanuel Lasker, New York 1924
  - Jose Raul Capablanca, New York 1927
  - Robert James Fischer vs Taimanov & Larsen 1971
  - Anatoli Karpov, Linares 1994
  - Garry Kasparov Linares 1999
  - Vladimir Kramnik vs Kasparov 2000
  - Magnus Carlsen, Nanjing 2009

Besides that, several correspondence games between the users at chessgames.com and various GM-s were analyzed to find out the absolute average error in circumstances where the quality of moves exceeds that of the analysis engine and hardware.

In order to find out the extent of the growth of playing strength in the course of time, moves from games grouped by decades from 1860s till 2000s were analyzed. Each decade had at least 250 positions valid for comparison. Games were chosen randomly from chessmetrics-rated 2600-2700 range. Thus we will have an idea which absolute level in terms of

today's rating chessmetrics rating 2650 corresponds to.

To find a correlation between modern rating and the accuracy of play, 9 cohorts at each 100 elo were analyzed in the range 1900-2700. In each cohort the rating range was [x-25 ; x+25], where x signifies the goal rating of a particular cohort. The lowest number of moves was 400.

And to find out how strongly engines play, 5 different chess engines from CCRL 40/40 rating list[3]. Taken into consideration were at least 250 moves by the following engines: Hiarcs 12.1 (2912), Crafty 23.0 (2630), Philou 2.8.0 (2367), Waxman 2008 (2116) ja Micro-Max 4.8 (1878). Ratings given as of september 2014.

The estimation of the accuracy of play is represented by the average gap in centipawns between a move suggested by the engine and the move made by a player. The engine was Rybka 3 MP, with 65 seconds per move. The chess interface was Arena 2. The hardware was Intel i7 860 @ 2.80 Ghz.

Only moves made in more-or-less even positions should generally be considered. If moves suggested by the engine and those actually made on the board were both outside the range [2.00 ; -2.00] and with the same sign, then a position was considered as decisive, and moves were discarded.

As a novelty, left out of consideration were moves that are very obvious. A move is considered as being too easy to spot if it meets the two criteria below starting from the first ply:

- a move suggested by the engine remains the same;
- the gap between the two best moves is always 1.00 or larger.

One must have in mind that there is a boundary above which the magnitude of errors is irrelevant. For example, if a player makes a move after which the evaluation drops from 1.23 to -3.04, and another move with the evaluation drop of from 0.89 to -11.03, then there's no basis for assuming that the former is objectively better than the latter. What actually matters is the fact that a good positions transformed into a lost position. It makes no difference whether a lost/won position is evaluated 2.00 or 15.00. The smaller the number of analyzed moves, the more such a phenomenon distorts results. For that reason the maximum boundary arbitrarily is set to 4.00.

Generally, knowledge of opening variations by heart is not counted among chess skills, although it's debatable. Since the opening theory is advancing vigorously each year, the start point of analysis is dependent on time periods:

1860-1879	1880-1899	1900-1919	1920-1939	1940-1959	1960-1979	1980-1999	2000s
8	9	10	11	12	13	14	15

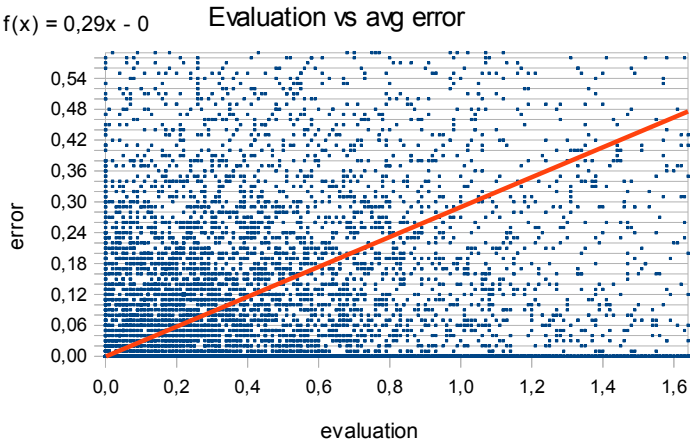
The minimum length of games is 20 moves + the start point as shown above. So the shortest allowed length of games varies between 28 and 35 moves.

Due to the evaluation functions of engines being mostly intended for middle games, the exact evaluation numbers are not so relevant in the endgame phase. Oftentimes they are not reliable. Therefore, in positions with less than 10 pieces all moves whose error is lower than 1.00 were discarded.

2.1 Impact of evaluation on the accuracy of play

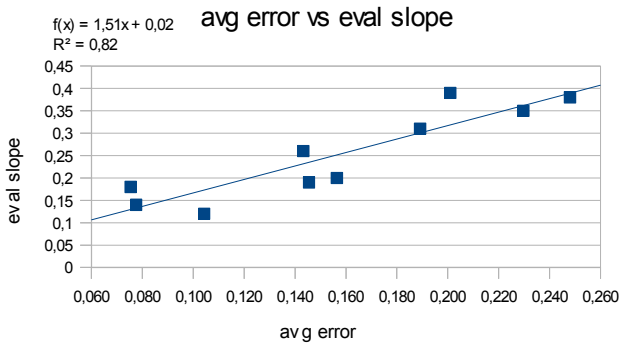
The skills of a chess player and conditions are not the only factors the absolute accuracy depends on. The magnitude of evaluations also has a role. It operates through a mechanism called the scaling effect. If one stretches out a rubber band, then all images on it would appear elongated. The same thing applies to engine evaluations. The higher evaluation numbers, the larger the gaps between moves become. Hence, before we can start observing the influence of various factors on the accuracy, it is necessary to find a hypothetical mean average error in the case of the evaluation being the same for all players.
The graph below shows a correlation between evaluation and the average error.

3 http://www.computerchess.org.uk/ccrl/4040/rating_list_all.html



Graph 1: evaluation vs average error

Each blue dot represents a position. The red line shows the linear correlation between evaluation and error, also called a slope. Unlike the factors of difficulty, the relationship between accuracy and evaluation does not depend on player's nature of play. Therefore, if the evaluation were the same for all players, the excpected error would have to be derived according to the same formula. But here a new problem arises: the average error varies among players, affecting the slope of the linear relationship. A slope indicates the degree of error change in relation to evaluation changes. To find the relationship between the slope and the average error, 10 randomly picked selections with 1500 positions in each were selected. The graph below shows the result that can be taken as a basis.



Graph 2: eval vs avg error slope depending on the average error

For example, if a player has the average error of 0.120, then, according to the formula, his slope would be 1.51*0.12+0.02 = 0.20. Increasing of the average evaluation by 0.1 would cause the player's average error to be inflated by 0.2*0.1 = 0.02.

2.2 Difficulty of positions

This research uses 2 different factors of difficuly:

- the difference between the best and the second best moves, expressed in centipawn units
- complexity

The first one is self-explanatory. The latter one needs some explaining. The manner of calculating complexity is taken from the work of Bratko and Guid. Every time the engine proposes a new #1 move, the gap between the best and and the second best moves is recorded, and at the end all these are summed together. Here it is presented in the form of original program code.

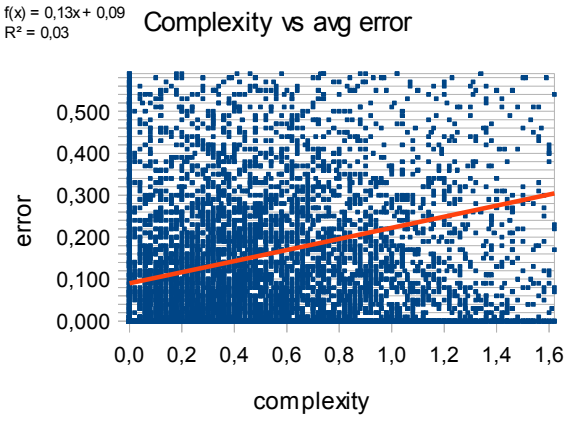
```
complexity := 0
FOR (depth 2 to 12)
IF (depth > 2) {
IF (previous_best_move NOT EQUAL current_best_move) {
complexity += |best_move_evaluation
- second_best_move_evaluation|
}
```

```
}
previous_best_move := current_best_move
}
```

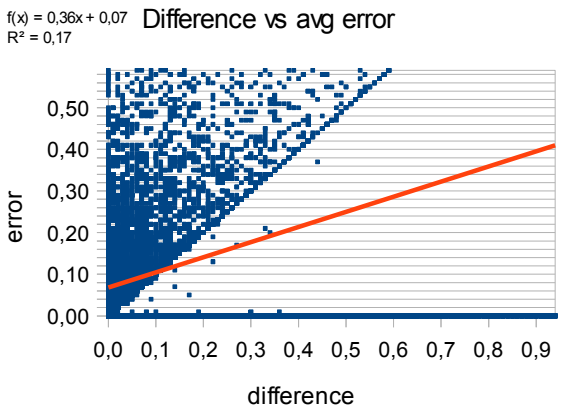
In this study a modified version is used. Through depths 10-15 plies all values are doubled to assign them more importance. It's always harder to see any changes in greater depths and indicates a more complicated position.
Below is a comparative example how computing complexity scores is carried out in both ways. Highlighted with yellow are cases where the best move changes. The sum at the lowest row shows the degree of complexity.

move	evaluation	difference	depth	move	evaluation	difference
Nc6	-0.23	0.00	2	Nc6	-0.23	0.00
Nf6	0.54	0.03	3	Nf6	0.54	0.03
Nf6	-0.21	0.00	4	Nf6	-0,21	0.00
d6	0.45	0.16	5	d6	0.45	0.16
Nc6	-0.10	0.00	6	Nc6	-0.10	0.00
Nc6	0.30	0.00	7	Nc6	0.30	0.00
Nf6	-0.03	0.02	8	Nf6	-0.03	0.02
Nf6	0.29	0.00	9	Nf6	0.29	0.00
Nf6	0.21	0.00	10	Nf6	0.21	0.00
Nf6	0.19	0.00	11	Nf6	0.19	0.00
Nf6	0.05	0.00	12	Nf6	0.05	0.00
e5	0.29	0.17	13	e5	0.29	0.17 (x2)
e5	0.21	0.04	14	e5	0.21	0.04
e5	0.26	0.14	15	e5	0.26	0.14
	sum	**0.38**			**sum**	**0.55**

The two graphs below illustrate how both factors of difficulty influence the accuracy of play. All analyzed positions are included.



Graph 3: complexity and average error



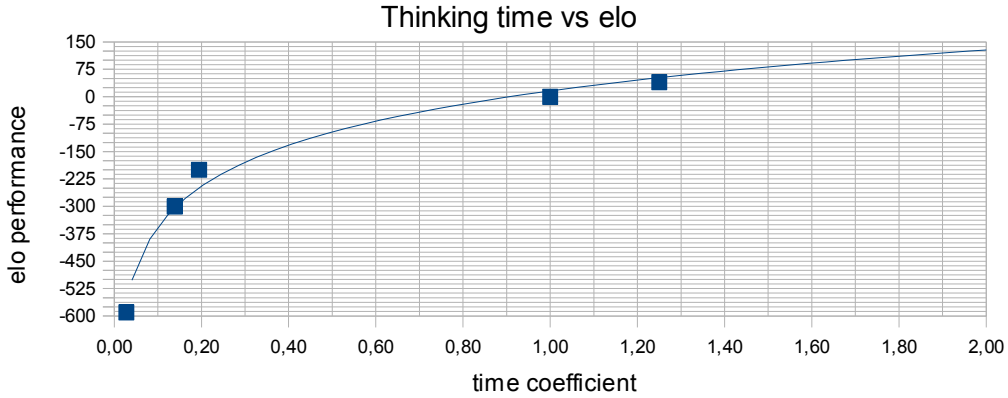
Graph 4: difference and average error

Influence of a factor of difficulty on a player is individual, and is dependent on one's nature of play. Some players are relatively more susceptible to changing difficulty of positions. Their accuracy becomes worse faster than other players with increasing difficulty. As the difficulty of positions cannot be described by one parameter only, it remains possible that different factors have different effect on a player. For example, in the instance of two equally strong players, one of them may have a lower than average tolerance for the factor represented by 'complexity' in this study, and a higher than average tolerance for 'difference'; but completely the other way around for another one.

2.3 Thinking time

There is plenty of information in the Internet about various time controls that have been used in various events throughout history. Unfortunately it is not always possible to find any information about in a certain event. In such cases the following principle was applied: 1880 - 1925 4 min; 1926 - 1945 3 min 20 s; 1946 - 1985 3 min 45 s; 1986 - ... 3 min per move.
K. W. Regan has found out that blitz (5'/game or 3' + 2"/move) - 575-600 elo; rapid 25' + 10"/move - 200 elo; rapid 15'

+ 10"/move - 300 elo.[4] According to that, the double difference in thinking time is equal to 112 elo, and the relationship between them is logarithmic. For engines the difference is worth 66 elo.



Graph 5: Dependence of performance on thinking time

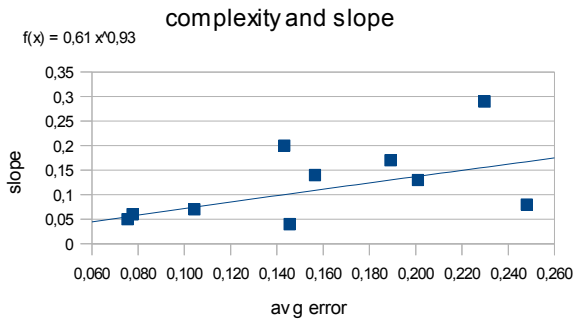
The biggest concern in games of earlier times is adjourned games. There's no doubt that a possibility to analyze games either alone or with assistants greatly helps the accuracy of moves played after resuming the game. It would be necessary to know how long those sessions lasted before resuming the play, and whether analyzing was allowed. As in the case of time controls, information is rather scarce. In the absence of reliable information, 1 hour was added to time control of each game that underwent adjournements as a compensation.
Sometimes the remaining number of moves after 40^{th}/60^{th} move has not been specified in time control information, except the number of minutes. In such cases the remaining time amount was divided by the number of moves actually played. If it exceeds time per move specified in the first part of time control, then the average thinking time in a given phase of the game is considered the same as in the preceding phase.

2.4 Practical play

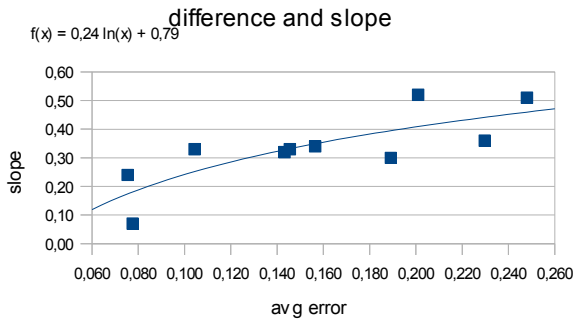
Of the three possible manifestations of practical play, only the difficulty of positions is looked at here. Ideally, it would have been preferable to use the suitability of position types and thinking time as well. But in the first case it would have been necessary to devise a way to quantify the suitability of the types of positions for players. In the latter case the knowledge on the precise amount of time spent on thinking on each move would have been needed. Both are unrealizable at the current juncture. The method in itself is simple – measure and compare the difficulty of positions for either side of the board. If one side has positions that are easier to play, it may be assumed that its results are better than its accuracy of play would suggest. The effect of difficulty difference between either player depends on two factors:

a) degree of the difference between the difficulty of positions
b) sensitivity of a player's accuracy of play to difficulty

Generally there's no data available on the tolerance of particular players with respect to changing difficulty level. In such cases it's possible to use generalized sensitivity to both factors of difficulty of positions and which is dependent on the average error. In order to find this, first we take the average rating of all opponents and look up its equivalent average expected error in the error–rating table. Secondly, we determine the relationship between average error and slopes for both types of difficulty, as shown in the graphs below. Similary to the graph 2, each data point represents a randomly-selected dataset of 1500 positions.



Graph 6: complexity slope depending on avg error



Graph 7: difference slope depending on avg error

4 http://www.chessgames.com/perl/chess.pl?tid=80980&kpage=20#reply535

Slope increases with average error. Hence, if our opponent had an average expected error of 0.205, its complexity vs average error slope would be $0.61*0.205^{0.93}=0.14x$ and difference vs average error slope $0.24*\ln(0.205)+0.79=0.41x$.
It's not necessary to include practical play if both sides of games are taken into analysis. In that case differences in difficulty, suitability, thinking time etc would cancel each other out. If a game for one player is on average more difficult by x hypothetical units, then his opponent has, at the same time, the game easier by -x units, and the sum would always be zero.
For this reason, practical play has only been included in the analysis of the games of the 7 most remarkable performances in the history of chess. As for the rest of games, both sides have been taken into account.

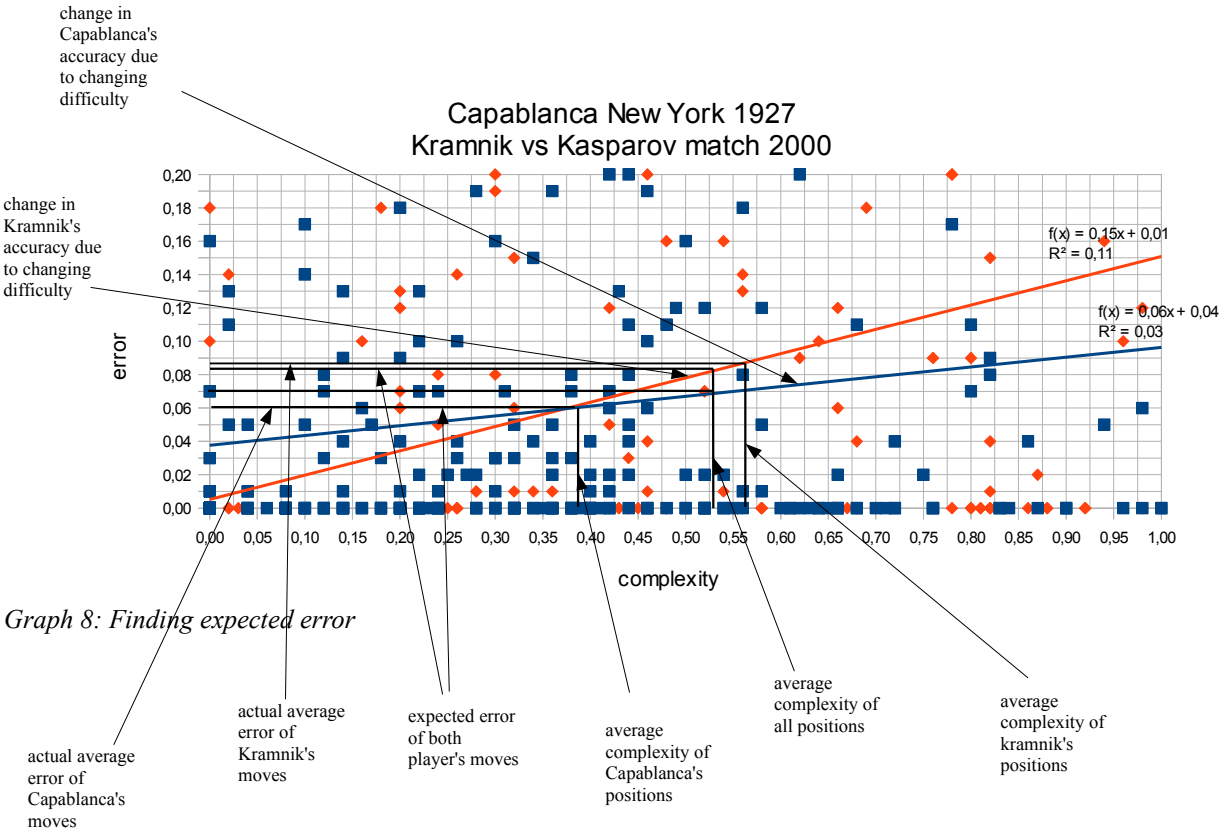
2.5 Finding the strength of play

Having determined the absolute accuracy of play and the aforementioned factors having effect on it, it becomes possible to derive the expected error of players. It consists of the following steps:

1. Find the average expected error of players.
2. Establish a relation between a modern rating and the expected error.
3. Find the modern rating equivalent of the expected error.
4. Find out rating losses/gains due to time control an practical play.

As a result we will get a supposed today's rating corresponding to the strength of play. Unfortunately, one must be satisfied with the fact that full confidence can never be attained. Methods described here are by no means 100% reliable, as it's still in its infancy and chess engines of today have limited abilities.

The expected error indicates a player's hypothetical accuracy of play (average error), if the difficulty of positions and evaluation were exactly the same for all players. In this study the average complexity of all moves valid for comparison – 0.528 and difference – 0.253 were used to represent a common ground. The graph below showing how the accuracy of Capablanca and Kramnik changes as a function of complexity also depicts the manner the expected error is determined with the help of linear trend lines.

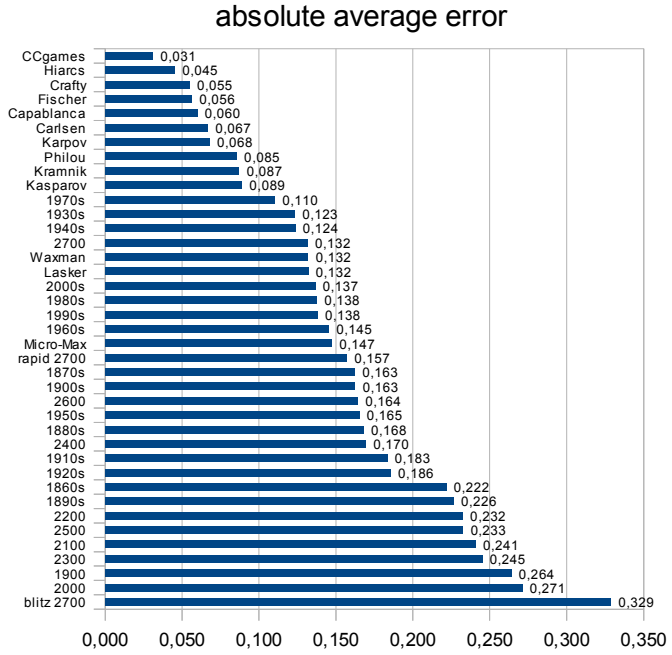


Graph 8: Finding expected error

As we can see, Capablanca's expected error by complexity is 0.070, and that of Kramnik is 0.084. Kramnik's positions were a little more complicated and those of Capablanca was far lower than the average complexity of all positions, therefore the gap between their accuracies of play would be smaller if they both had positions of the same complexity. The expected error according to the difference is found by the same method. One can also note that Capablanca's accuracy of play has been less dependent on difficulty than Kramnik.

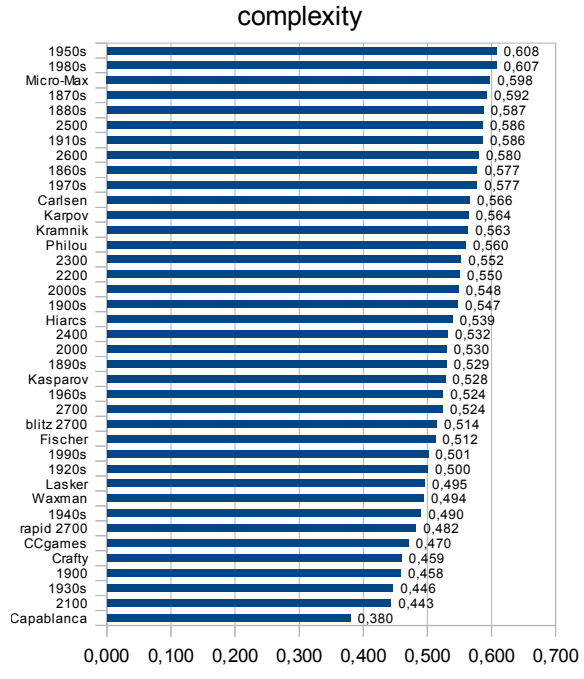
3. Results

The following section is divided in two parts. First all necessary data on all analyzed chess-playing entities will be dealt with, and then, step-by-step based on that, we'll find the hypothetical strength of play of each player.
The most important of these is, of course, the actual accuracy of play i.e. the average error. The result of a game only depends on differences in the accuracy of play. However, it must be born in mind that it never directly shows the level of chess skills, but rather remains biased towards players with more positional style and longer time controls. The following graph displays all chess-playing entities sorted by average error.

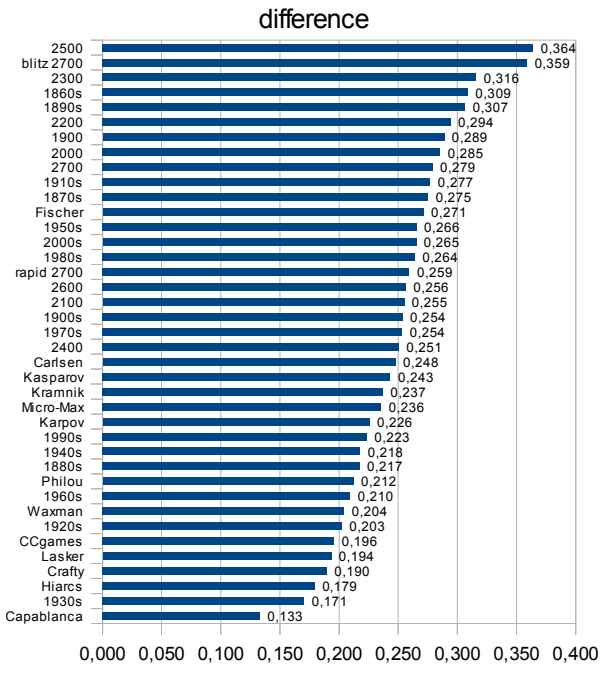


Graph 9: Absolute average error of moves

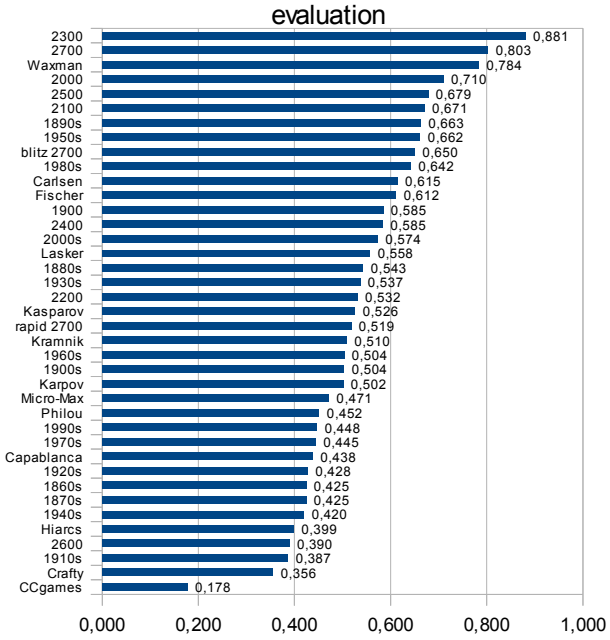
Expectedly, most engines occupy top spots. We also can see that generally today's players and those with higher rating are placed higher. The best result was achieved by correspondence games of chessgames.com. Since in all positions in those games the quality of play exceeded that of analysis, the figure 0.031 does not represent the actual level, but the boundary of trustability.
The 3 graphs below rank players according to the both factors of difficulty and the average evaluation.



Graph 10: Average complexity of positions

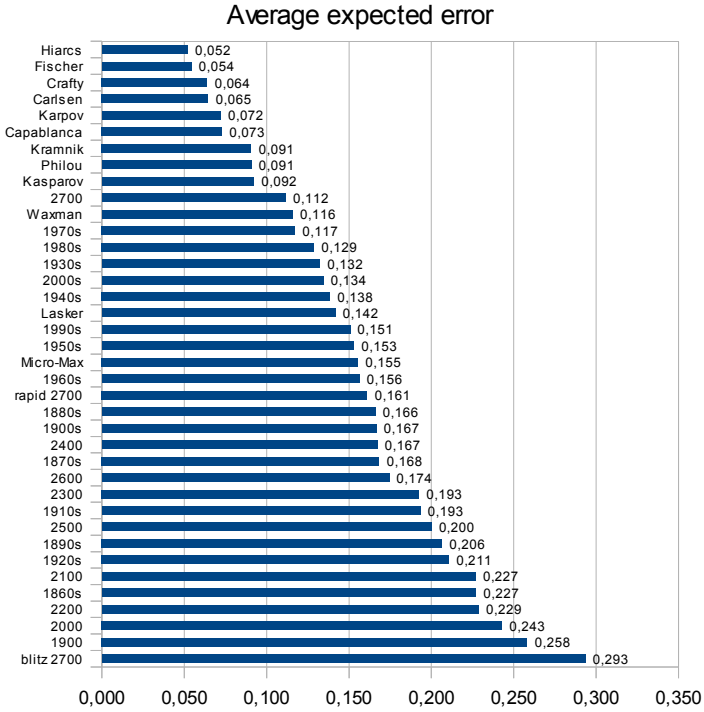


Graph 11: Average difference of positions



Graph 12: Average evaluation of positions

It stands out that Capablanca had positions with by far least difficulty. There has been a lot of mentioning on Fischer's simple style of play, his tendency to avoid complications. Indeed, according to the graph 11, the complexity of his positions were below average in games against Larsen and Taimanov. However, it can be seen on the gaph 10 that the average difference between two best moves in Fischer's positions is above average. The fact that a position seems somewhat easy to us does not automatically mean it would be easy to find accurate moves. It is perhaps not quite surprising that correspondence games from chessgames.com have the lowest average evaluation, i. e. in those games the positions were equal longer due to higher quality of play.

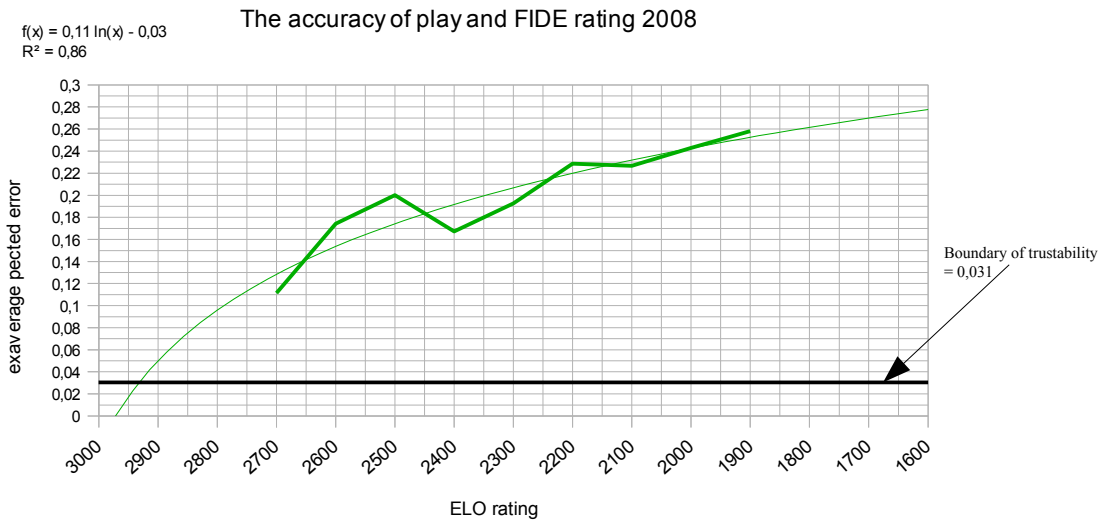


Graph 13: Average expected error

The graph above shows the average expected error which is derived by taking the average of both expected errors by complexity and difference and includes changes in the average error due to the evaluation. The results are more logical, compared to what was dispalyed on the graph 9. Correspondence games are left out, as there is no point in measuring changes in the accuracy of play, if its estimation cannot be trusted. As a rule in all kinds of measurements, the gauge

must be of higher quality or trustworthiness than things being measured. The methods used in this paper are simply not adequate enough for modern software-assisted correspondence games.

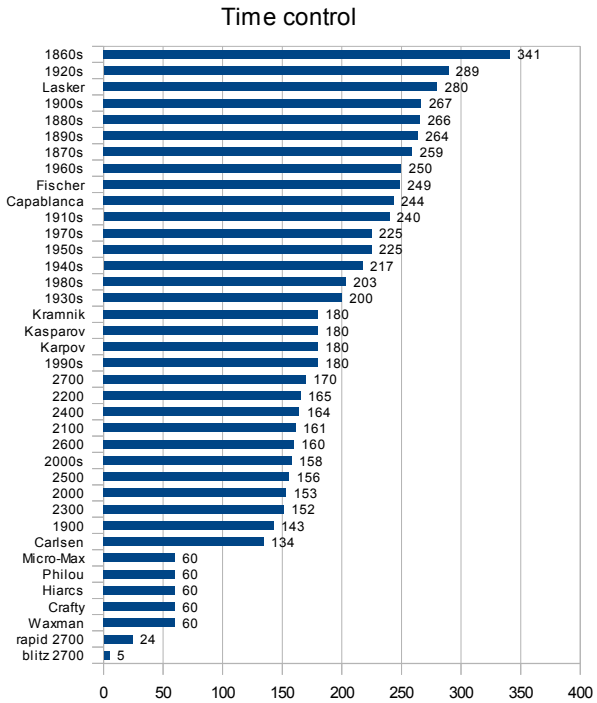
The next step is to take data from the previous graph to find the relationship between the rating and the quality of play.



Graph 14: The relationship between accuracy and fide rating 2008

The relationship appears to be logarithmic. The black line depicts the approximate boundary of trustability below which engine output cannot be trusted. It is interesting to note that it crosses the trend line at 2931 ELO, which may indicate that the level of play of the combination of the engine, hardware and time used here is equal to 2931 FIDE 2008. But that is naturally a speculation which needs further research.

Players ranked according to thinking time:

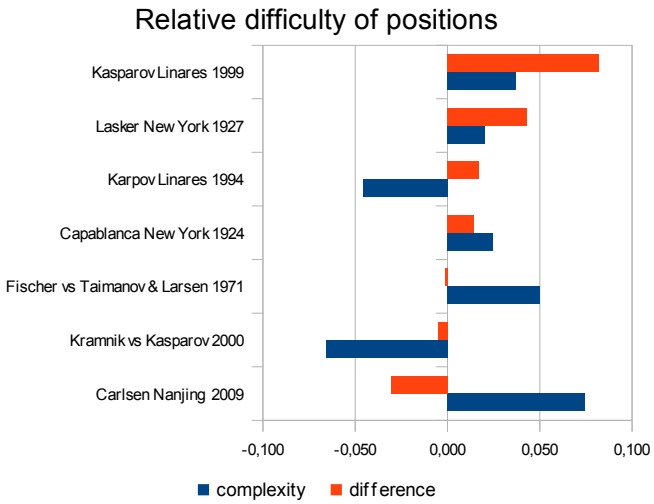


Graph 15: Thinking time

The farther back in time, the longer time controls are.

The next steps represent an attempt to factor in at least a fraction of generally unfathomable and messy notion called practical play.

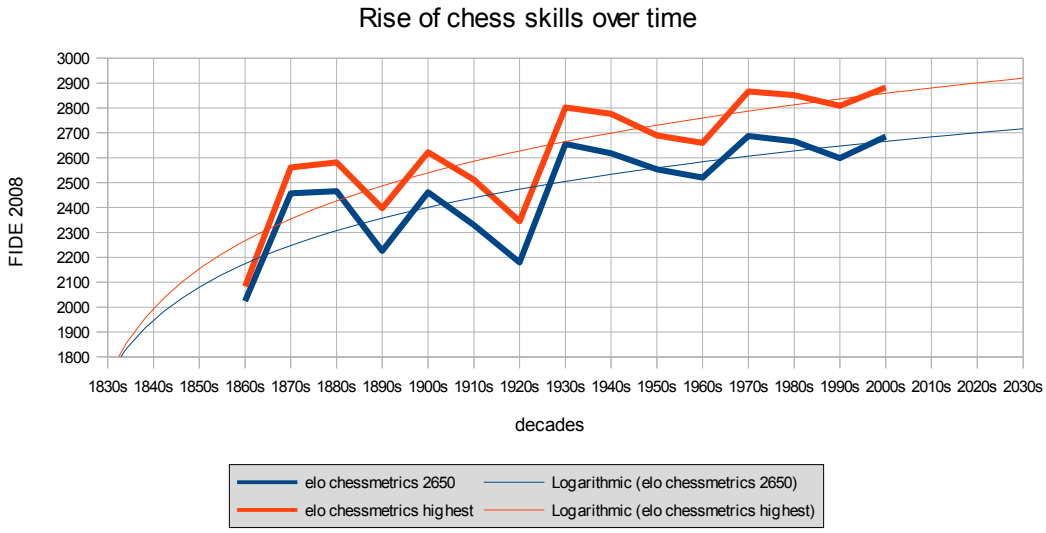
Players ranked according to relative difference of positions:



Graph 16: Relative difficulty of positions

Negative value shows that opponents had positions easier, in the case of positive one it is the other way around. As one could have expected, Kasparov and Lasker, players known for practical play, are situated on top. Somewhat surpisingly, it appears that even Capablanca too had both difficulty factors easier than his opponents. One of reasons could be that the easier one's positions, the greater the probability that the opponent's positions are more difficult, despite the degree practicality in one's play. Kramnik's positions were expectedly easier than Kasparov's in the title match 2000.

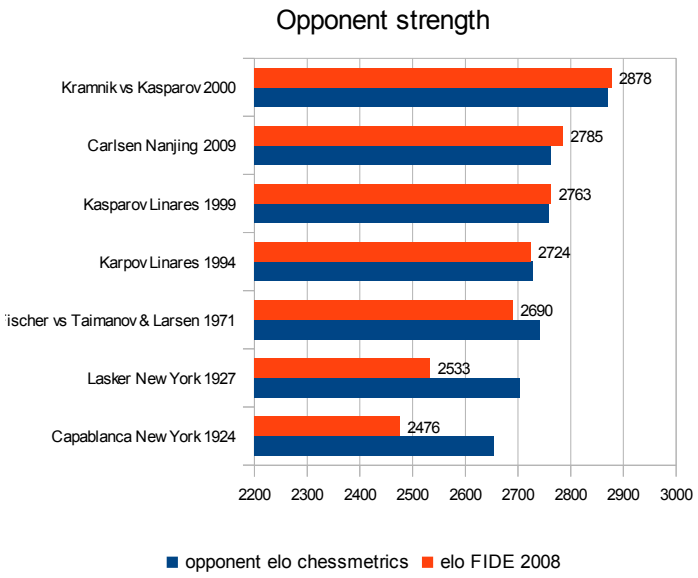
Before trying to find out how much exactly a difficulty differential influences opponent's play, it is necessary to know the strength of opponents. First we look up FIDE or chessmetrics rating of that time and translate it into contemporary rating equvalent.



Graph 17: evolution of chess strength by decades

The blue line represents actual data based on the analysis of randomly picked games (rating range 2600-2700) from each decade. The red line represents top-rated players' strength of play. The gap in each decade between a top-rated player and a 2650-rated player is based on the arithmetical averages of january lists in the same decade. It can be seen that if the logarithmic trend line can be trusted, the first time top players reached an FM level (2300 ELO) already in mid-19th century. The level of an International Master (2400-2500) was achieved in 1880-1890s. GM level was reached during the first decades of the XX century. Development was relatively quick-paced at that time. Top players were on today's Super GM level already in the 40-ies. With time, because it's always incresingly harder to make progress the closer one gets to perfection, the tempo of improvement has continuously decreased. The graph shows that players supposedly will reach the level of play which in the year 2008 would have been enough for 3000 ELO in the final decades of the current century. But extrapolating that far must be taken with grain of salt. It also probably is not correct that in the beginning of the 19th century they played at 1500-1600 level.

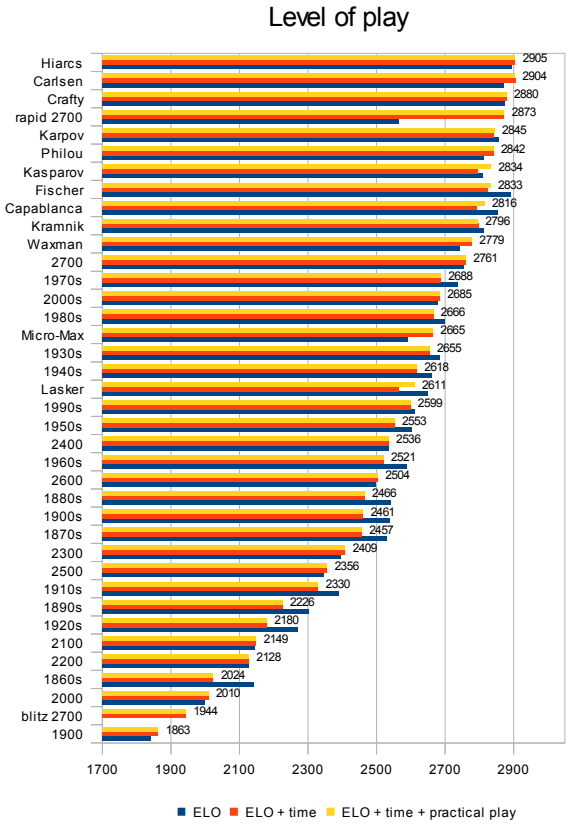
The graph below demonstrates the opponent ratings and their actual strengths.



Graph 18: Opponent ratings and strength of play

Not surprisingly, Kramnik had the strongest opponent against Kasparov in 2000. The weakest opposition was against Capablanca in New York 1924. By translating those ratings into average expected error and taking into account how generalized sensitivity, as described in the section 2.4., to either type of difficulty depends on it, it can be ascertained how much difficulty differentials affect performance.

The next graph shows the final conclusion of this work. The blue bars indicate ratings directly derived from the expected error, as shown on the graph 14. Whereas the red and yellow bars respectively indicate ratings after including influences from time controls and practical play on the performance.



Graph 19: Level of play of various chessplaying entities in terms of ELO 2008

The winners according to this criterion are Hiarcs 12.1 and Carlsen in Nanjing 2009. It may seem surprising that his actual accuracy of play in that tournament was almost 100 points lower than his official TPR (3001). But it can be explained by two facts: against Wang Yue in round 8 he made 2 blunders, which is too much for a player like him in a won game that was not overly complicated; Nanjing 2009 was a relatively short tournament - 10 rounds only, increasing the likelihood of large discrepancies between actual level and TPRs. It is hardly a surprise that at the bottom of the graph are situated those who should be there – 1900-rated players and blitz games of 2700-rated players.

4. Miscellaneous

In this section several additional interesting conclusions that the collected data offers will be provided.

4.1 Chessmetrics 3-year peak top 50

The table below compares 3-year peak ratings for each player taken from the chessmetrics.com site and their FIDE equivalents in 2008. The year indicates the middle year of the three-year periods.

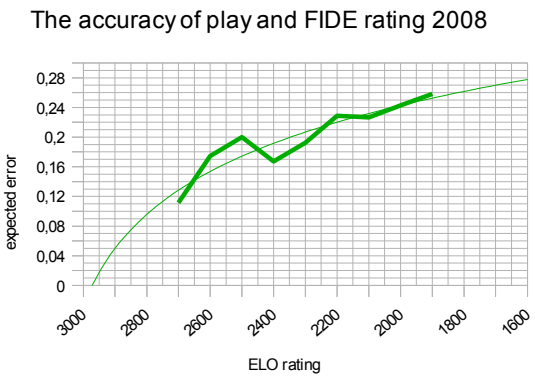
	name	year	chessmetrics	FIDE 2008
1	Kasparov	1990	2874	2861
2	Fischer	1972	2867	2817
3	Capablanca	1920	2857	2664
4	Lasker	1895	2855	2562
5	Botvinnik	1946	2852	2738
6	Alekhine	1931	2841	2684
7	Karpov	1989	2833	2819
8	Anand	1998	2822	2825
9	Kramnik	2001	2815	2824
10	Pillsbury	1901	2806	2540
11	Maroczy	1906	2799	2554
12	Korchnoi	1979	2798	2763
13	Tarrasch	1895	2796	2503
14	Ivanchuk	1992	2794	2785
15	Steinitz	1885	2794	2451
16	Smyslov	1955	2793	2703
17	Petrosian	1962	2789	2716
18	Tal	1960	2786	2708
19	Rubinstein	1912	2781	2559
20	Reshevsky	1953	2776	2681
21	Najdorf	1947	2775	2664
22	Zukertort	1884	2774	2425
23	Keres	1956	2773	2685
24	Nimzowitsch	1929	2770	2607
25	Bronstein	1951	2770	2669
26	Spassky	1970	2767	2712
27	Kamsky	1995	2765	2762
28	Chigorin	1896	2763	2475
29	Marshall	1917	2759	2556
30	Leko	2001	2757	2766
31	Janowsky	1904	2757	2504
32	Fine	1940	2756	2626
33	Topalov	1997	2754	2755
34	Salov	1994	2754	2749
35	Gelfand	1992	2754	2745
36	Shirov	2000	2753	2760
37	Bogoljubow	1927	2753	2583
38	Geller	1963	2752	2681
39	Morozevich	2000	2751	2758
40	Euwe	1936	2750	2608
41	Adams	2001	2749	2758
42	Polugaevsky	1977	2748	2709
43	Beliavsky	1988	2747	2731
44	Timman	1989	2747	2733
45	Schlechter	1911	2747	2522
46	Portisch	1980	2746	2713
47	Stein	1966	2745	2681
48	Vaganian	1985	2744	2721
49	Jussupow	1987	2744	2725
50	Larsen	1970	2744	2689

According to that table, the strongest level of play of all times was performed by Kasparov during 1989-1991 where his play supposedly would have been rated circa 2860 in 2008. Yet, it must be taken into account that this table is a bit

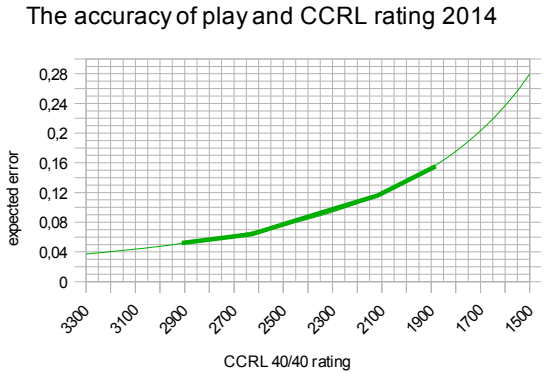
misleading. It is often so that the chessmetrics rating of a player is a decade or more later only a few points below his peak, but nevertheless his play is better due to general rise in chess skills. For example, Lasker's chessmetrics rating in 1894 was 2878, but in 1917 it was 2860, whose 2008 equivalent would have been ca 2650. Kasparov's rating in 1999 was 2884, merely 2p lower than his best he achieved in 1993, but there is no doubt that his actual quality of play had improved by that time.

4.2 Comparison of human and engine ratings

In this work there is enough data on both humans and engines for making interesting comparisons between so different types of players. Below is a side by side comparison of the relationships between the accuracy of play and ratings of either type. CCRL ratings are given as of 17.10.2014. According to the site, the time control was chosen in such a way as to be equivalent of 40 moves per 40 minutes on Athlon 64 X2 4600+ (2.4 Ghz).



Graph 20: FIDE rating vs accuracy

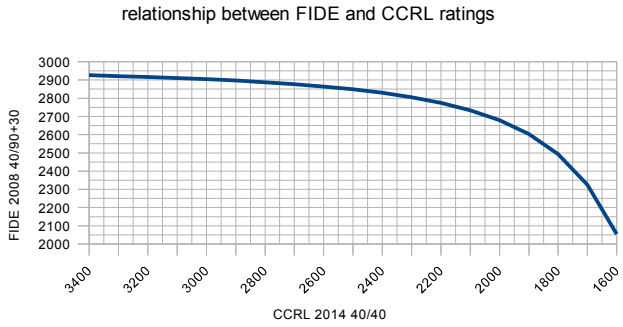


Graph 21: CCRL rating vs accuracy

As we can observe, the trend lines are of opposite nature. The relationship between the accuracy of play of human chess players and the rating is logarithmic; on lower levels, the accuracy gaps are smaller than at the top. On the other hand, in the case of engines, it is completely opposite – exponential. It should be noticed how closely the trend line follows the actual line representing the accuracy of play; there is a stark contrast. It confirms what was known for long – computers' play is far more stable.

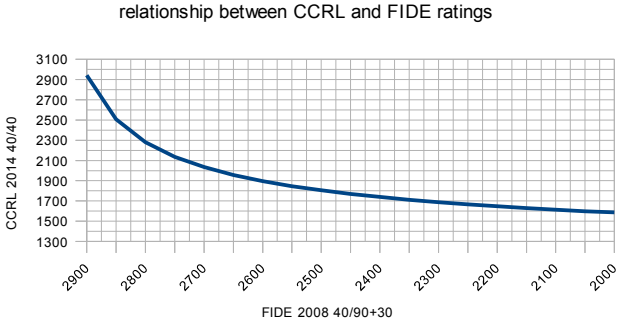
Based on data on those two graphs, it is possible to compile conversion tables for finding one-on-one correspondences between both rating systems.

CCRL 40/40	FIDE 2008 40/90+30
3400	2926
3300	2921
3200	2916
3100	2911
3000	2904
2900	2897
2800	2888
2700	2877
2600	2864
2500	2849
2400	2830
2300	2805
2200	2774
2100	2734
2000	2679
1900	2603
1800	2493
1700	2325
1600	2054



Graph 22: human and engine rating comparison

FIDE 2008 40/90+30	CCRL 40/40
2900	2941
2850	2507
2800	2281
2750	2136
2700	2034
2650	1957
2600	1896
2550	1847
2500	1805
2450	1770
2400	1739
2350	1712
2300	1688
2250	1667
2200	1647
2150	1630
2100	1614
2050	1599
2000	1586



Graph 23: engine and human rating comparison

At first sight, it may seem surprising that the best chess engines are, according to this, so weak compared to humans. But, it must be taken into consideration that CCRL games are run on a quite weak hardware and the rate of play is nearly 3x quicker than the standard FIDE time control. It can be concluded from the graph 22 that in the beginning it was quite easy for engines to make progress against humans, but with time it is getting increasingly harder. Note: comparisons were made on the assumption that humans play against engines as they would against other humans; i. e. not using any anti-computer strategies. Unfortunately there is not yet a reliable way to emulate anti-computer play and its effects.

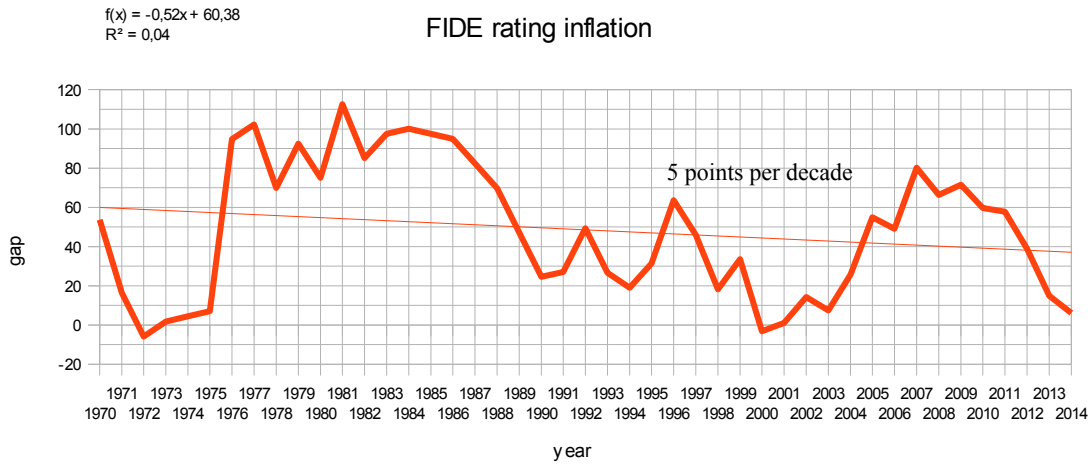
The reasons why the relationships between the accuracy and strength of play are exactly like that, are unknown to the author. One of feasible reasons could be that the nature of the curve is related to the relative importance of calculation-evaluation. The larger the relative importance of calculation in the move-choosing process (engines), the steeper the exponential curve is; while gaining in rating points, there is an ever-decreasing rate in the accuracy gain. But, if a player has a larger relative importance of evaluation (humans), then there is a contrary phenomenon: rating growth means a faster increase in the accuracy. If it is true, then there presumably must be such a hypothetical mutual relationship between calculation and evaluation where the accuracy vs rating relationship is linear.

4.3 Rating inflation and deflation

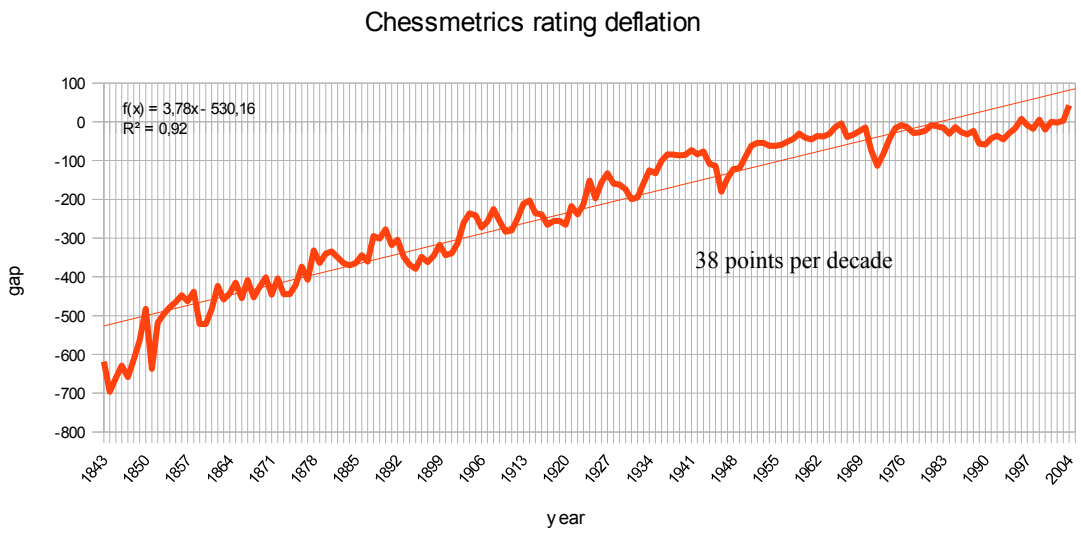
Whenever one deals with various rating systems, it may be assured that he finds a phenomenon called rating inflation which has lately been in the hotspot of intense debate. It is also the reason why it is pointless to compare players from different time periods based on ratings. Rating inflation has 2 different definitions: 1) inflation with respect to playing strength; 2) inflation with respect to the number of players.

The first type of inflation entails rating growth which is faster than increase in strength of play. In the case of the second type, the driving force behind that is the growth of the rating pool. The latter type has 2 different indications: 1) increase in the number of players above a certain rating threshold; 2) Increase of the rating number of the first-ranked player. The latter phenomenon is also existent at the first type of inflation.

Negative inflation, that is deflation, only occurs when the situation is the other way around; the rating rises slower than standard of play, the number of players wanes, or the rating of the first player decreases. There are a lot of attempts made to explain the existence of the inflation. Most popular among them are the incresing numbe rof players and lowering of rating floors. Here only the first type – skill-related inflation is under consideration. The two graphs respectively display FIDE rating inflation and Chessmetrics rating deflation. The gap is the difference between the 2008 FIDE playing strength equivalent of the best player, as represented by the red line on the graph 17, and the actual rating at that time.



Graph 24: inflation in the FIDE rating system



Graph 25: deflation in the chessmetrics rating system

On the first graph we can see that the gap has decreased from 60 points in 1970 to 37 points in 2014; 5 points per decade. Clearly recognizable are two 'mountains' and four 'valleys'. The valleys refer to periods where rating numbers were quite high because of a dominant player – Fischer, Kasparov (two periods), and Carlsen. The mountains mark periods with equality and no clear dominator. On the other graph there is a completely different situation. The ratings of 1st players of the rating lists have been relatively stable since 1890s, while the skill level has steadily been rising. In other words – what we see there is the deflation in the chessmetrics rating system. The rate of deflation has decreased somewhat since 1960s, which is logical, as the rate of improvement of playing skills must slacken over time. Here too, 'valleys' from domination periods can be seen. The rate of deflation is 608 points wihin 1843-2004 – 3.78 points per year.

4.4 Various trends

Previously we looked at how the strength of play had changed across the history and the two distinct rating systems. However, the same can be applied to changes in other factors, such as slope and both factors of difficulty. Before having a closer look at those, a short introduction on the notion of 'slope' and what it actually indicates will be presented below.

As persons more familiar with chess know, chess players can be split into two groups based on the nature of play:

1 positional, where intuition and knowledge prevail
2 tactical, where the speed of calculations, precision and creativity are most important

Usually it is known that nature of play dictates the choice of openings and the type of positions, but differences are also present in players' tolerances with respect to the difficulty of positions and thinking time. If one tries to solve a problem by calculating variations and possible outcomes, then it generally takes a lot of time before a solution is reached. On the other hand, it is universal: calculation is suitable for solving any type and however difficult problems. The advantage of problem solutions based on knowledge or intuition is speed. It takes almost no time to recall facts in memory or realize something via intuition. Their disadvantage is the fact that it is only suitable for relatively simple and more familiar problems; in case of solutions being illogical and unexpected, it fails. From this, the following facts follow:

1 players of positional type are relatively less sensitive to thinking time, but more sensitive to the difficulty of positions
2 it is contrary with tactical players: they are less sensitive to the difficulty of positions, but more sensitive to thinking time

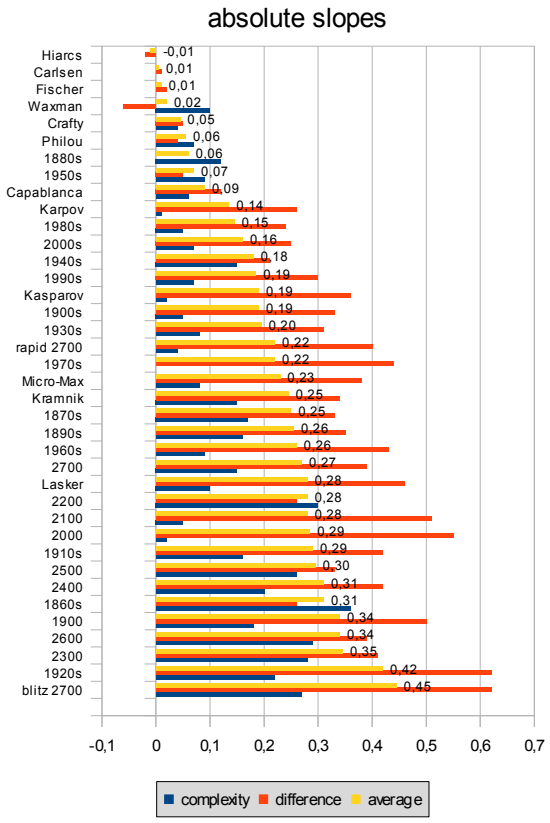
Hence the fact that the size of the slope of the relationship between average error and a factor of difficulty depends on player type. Tactical players have it smaller, positional ones bigger. Such a phenomenon may give us a simple method to find out which players have bigger relative importance of calculations and which ones intuition/knowledge in their move-finding processes.

The graph on the left shows the absolute average slope of all entities covered in this study. But, as it can be seen on the graphs 6 and 7, the size of slopes is dependent on the average error. Therefore, it is more preferable to determine how

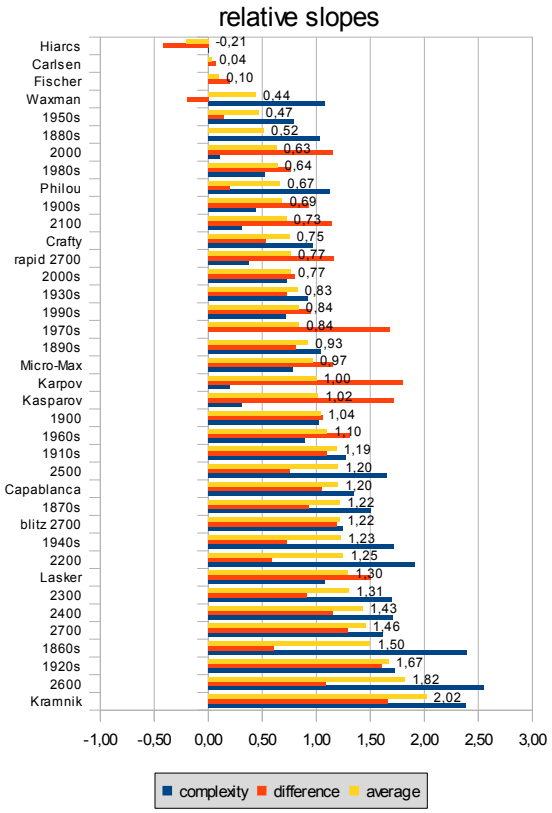
much the actual slope deviates from the expected slope. The formulas for calculating the expected slope are given here:

$$complexity\ relative\ slope = \frac{absolute\ slope}{0.61 * avg\ error^{0.93}}$$ and

$$difference\ relative\ slope = \frac{absolute\ slope}{0.24 * \ln(avg\ error) + 0.79}$$

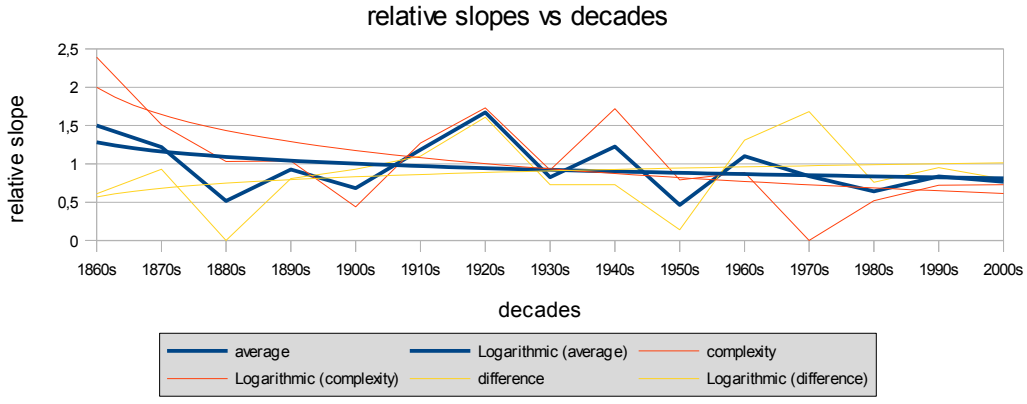


Graph 26: players sorted by absolute slopes

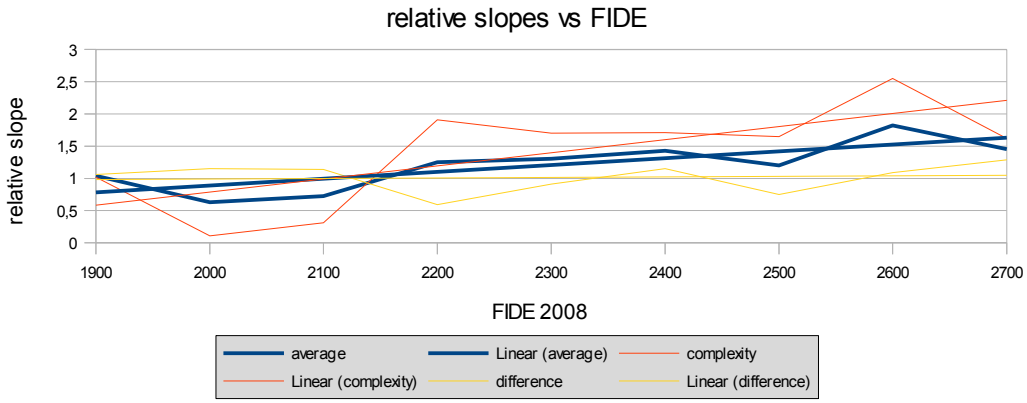


Graph 27: players sorted by relative slopes

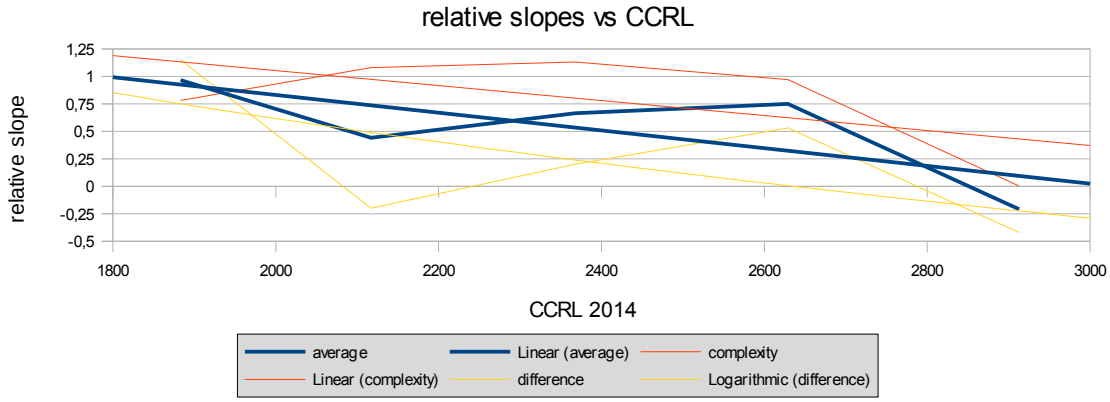
The smaller the digit, the bigger is the importance of calculations. The graph on the right reveals that all chess engines are at the top half, according to expectation. Capablanca and Kramnik are situated at the bottom half, confirming the common belief that those players were primarily intuitive players. Perhaps surprisingly, Kasparov and Karpov stand close to each other and Lasker so far down.
The changes of the average relative slopes across time and both rating systems are presented below.



Graph 28: relative slope across time periods



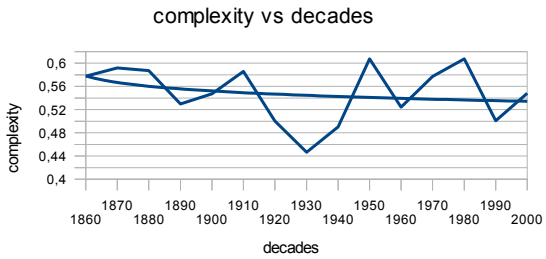
Graph 29: relative slope across FIDE rating



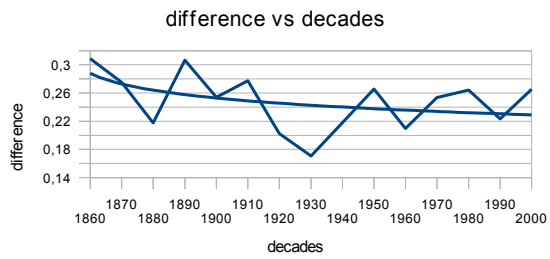
Graph 30: relative slope across CCRL rating

It appears that the average relative slope slowly decreases with time; the rate of decrease is even bigger in the CCRL rating system. It exhibits a completely different behaviour in the FIDE rating system where stronger players have bigger slopes. In other words, today's chess players have become more calculative that they were in the past. In a sense, it is logical; after My System by Nimzowitsch there have been no significant breakthrough in chess middlegame theory. While earlier the secrets of positional play were known to top masters only, nowadays, in the era of the worldwide spread of internet, chess books, engines etc, even mediocre players have a decent knowledge on chess. It means it is becoming increasingly more important to know how to play in dynamic and complicated positions where success depends more on calculations. The same phenomenon appears in the CCRL rating system. Here it seems that stronger chess engines have bigger dependence on tactical play. It may hint at the fact that search function is more important than evaluation function. As most people agree, in chess tactics always prevails over strategy, and better engines are more effective and quicker in finding good positions and illogical moves. Computers are good at doing calculation-based tasks, but given the enormous size of the search tree, this fact alone is of no usefulness, a programmer still has to carefully guide his program through variations. The case of FIDE ratings is completely different, higher standard of play also means higher importance of knowledge and intuition. How could it possibly be true, taking into account previous satements? A most likely explanation would be that human brains are very weakly suited for calculation. This ability can be trained a bit, but nevertheless it still cannot jump over its narrow limits. Knowledge, on the contrary, can be expanded a lot. Tactics training is more about improving pattern recognizing than pure calculation skills.

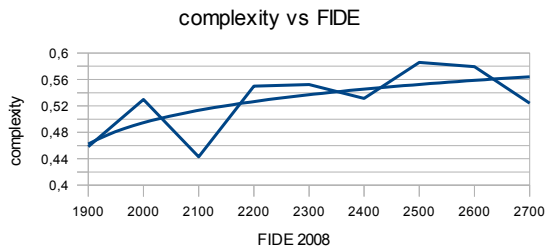
The following graphs display how difficulty of positions has changed.



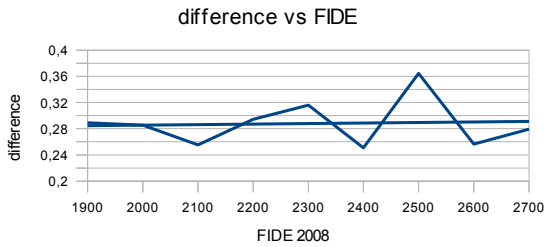
Graph 32: complexity across time



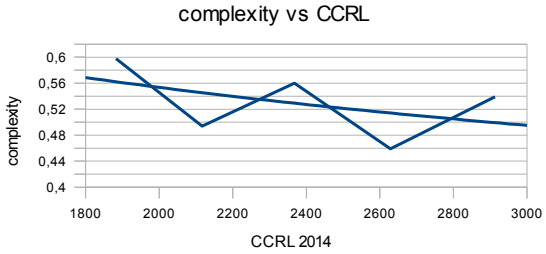
Graph 31: difference across time



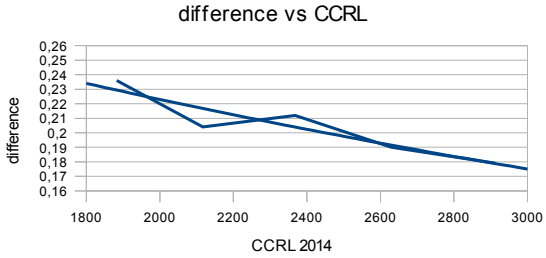
Graph 33: complexity across FIDE rating



Graph 34: difference across FIDE rating



Graph 35: complexity across CCRL rating

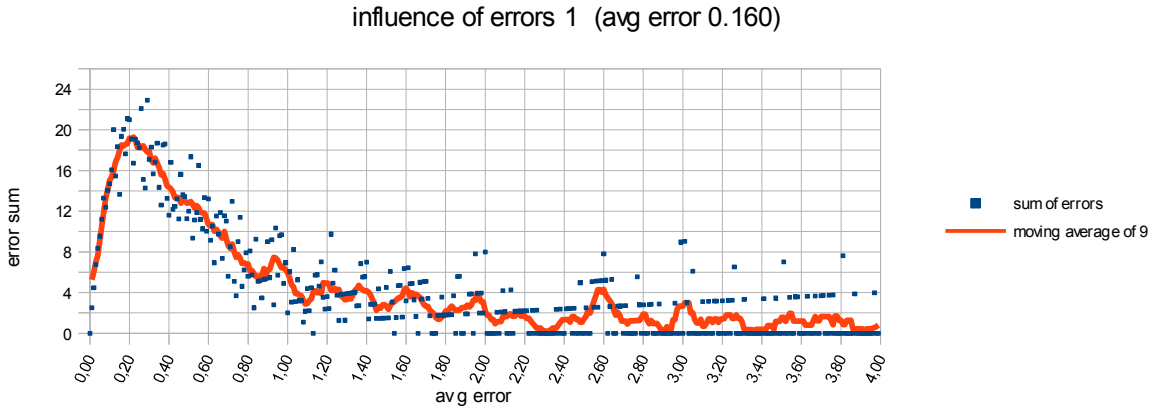


Graph 36: difference across CCRL rating

Positions have become somewhat easier compared to earlier times. Especially eye-catching is the low point between 1920-1940. Regarding FIDE rating, it looks like stronger players have a tendency to make positions more complicated. Better chess engines, on the other hand, end up playing in relatively easier positions.

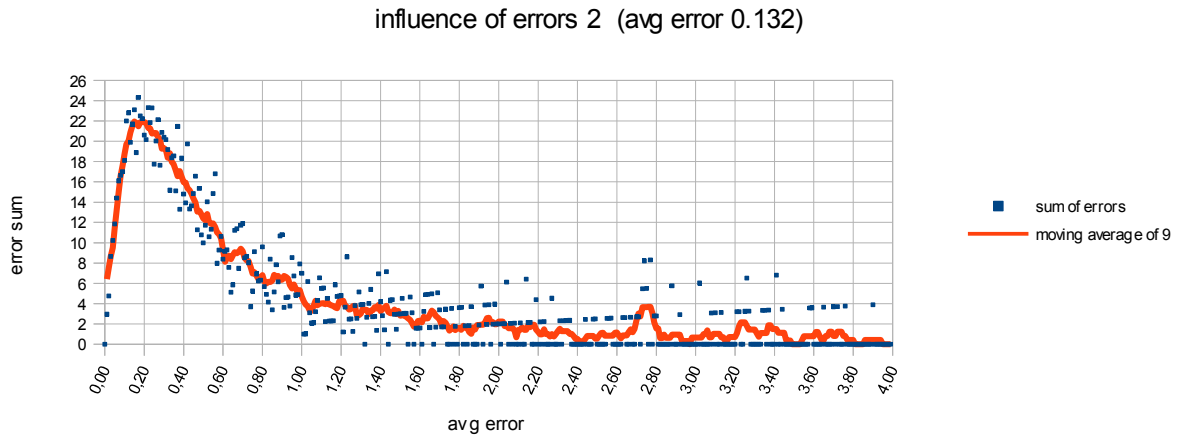
4.5 Influence of errors

And as a final part, here is data on the influences of errors of various magnitude. The influence of errors is calculated by multiplying the frequency by its magnitude. By comparing the resulting number with those of other errors, it can be ascertained which magnitudes of errors are the biggest source of inaccurate play. On the graphs below each blue datapoint marks the product of the magnitude of an error and frequency. The red line shows the moving average of 9 datapoints. Upper graph is based on data used in this study (12777 positions, average error 0.160), lower graph represents data taken from an earlier study[5] (14 174 positions, average error 0.132).



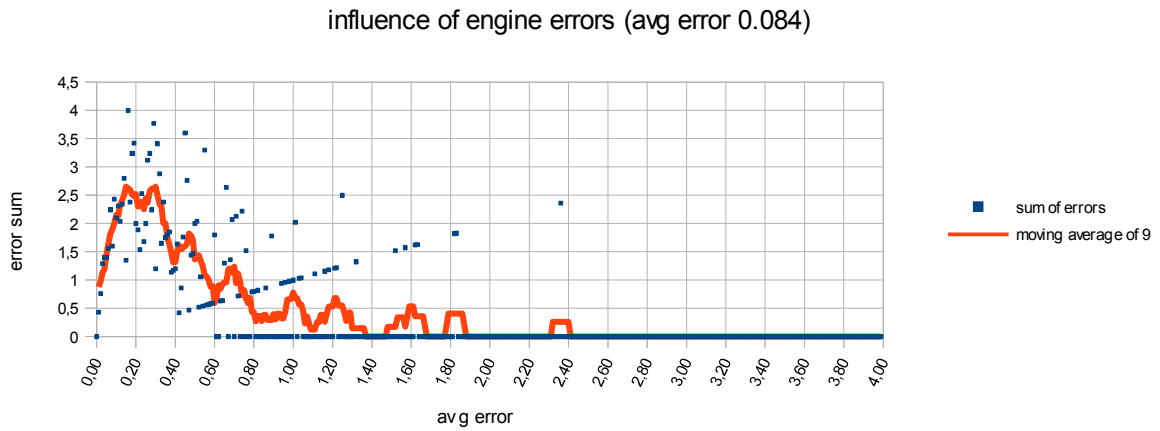
Graph 37: influence of errors 1

5 http://www.chessanalysis.ee/summary450.pdf

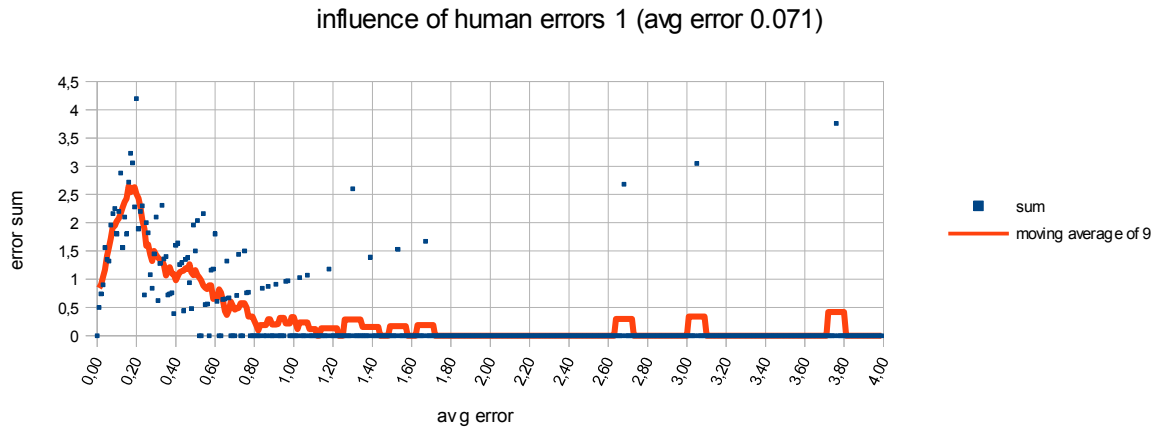


Graph 38: influence of errors 2

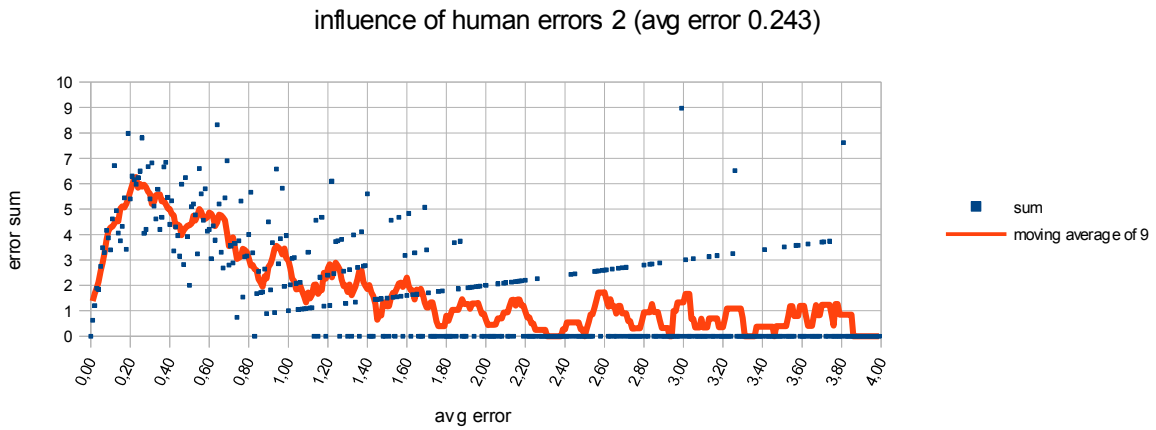
Despite of the fact that the average error is significantly different in both dataset, side-by-side comparison reveals, as shown by the red moving average line, biggest influence is roughly equal, both are around 0.20. Hence the question: will the main source of inaccurate play also remain the same if we have a separate look at engines and humans with roughy same accuracy? On the following graphs, the overall average error of all engine moves is 0.084. Human moves were grouped in two, one based on players with average error higher than 0.100 and those whose average error was lower than 0.200. The average errors of all moves combined were 0.071 and 0.243 respectively.



Graph 39: Influence of engine errors



Graph 40: Influence of human errors 1



Graph 41: influence of human errors 2

The results might be described as remarkable. All three graphs show that, irrespective of the accuracy of the moves, the error influence peak is persistently situated around 0.20 mark. True, in the case of engines, there is a small oddity, the peak seems to have a little cavity centered around 0.20, with two ridges surrounding it. It is presently unknown what may be the cause for that. These graphs also explain why it is not recommendable to use threshold-based analysis, at least not with the threshold values of 0.10 and above. At first glance, small to medium-sized errors may seem insignificant, but what they lack in gravity, they make up for being more numerous.

5. Conclusion and future perspectives

In this study the author, using Rybka 3, tried to measure the objective strength of play and to determine its relationship between wither type of rating systems. Besides that the aim was to record the change of the strength of play through time. In could be compared to athletics world record progression or world leading mark tables, that provide a good overview of development in athletics. Unlike many sports today, chess has been played for centuries, and the level of play has been since long ago been very high. As it became clear earlier, already by the beginning of the previous century, top players were on a par with today's weaker GM-s. In the light of this info, John Nunn's speculation that Hugo Süchting at Karlsbad tournament in 1911 was merely a 2100-rated player, should be regarded as a serious underestimation. According to chessmetrics rating after the tournament, he was only 253 points short of Lasker, which would rate the latter ca 250 points lower than can be seen on the graph 17. Due to the fact that so far there has been no reliable method for measuring the quality of play, such a phenomenon as overhyping of players of the past has gained ground. Surprisingly many people seriously think that former great figures were at least as talented as today's top players, and that they played as well or even better. Psychologically completely understandable, practically unnecessary; we all tend to see the past more beautiful than it actually was. Players of the past are being overrated also because out of all their games, there is a tendency to selectively highlight better specimens, whereas in the case of contemporary players, various sites providing live engine-assisted analysis display their average level. And since the population of the world and the number of chess players back then were smaller, the same thing can be said about talent pool. It is more probable to find more naturally talented players in larger pool.

Comparison between CCRL and FIDE rating systems gave a surprising conclusion. Before that, the author held an opinion that both systems had an analogous relationship between strength and accuracy. It comes out that the relationships are of opposite nature. The accuracy of humans decreases logarithmically with strength of play, with gradually diminishing rate, but the accuracy of play of engines, on the other hand, decreases at exponential rate. The fact that engines from the bottom part of the rating list are weak has been noted already long ago. One conclusion that can be made is that it is virtually impossible to reach negative ratings in engine rating systems, whereas it is very easy in the FIDE rating system. The weak point in the conclusion is that, because of the lack of proper methods, it was not possible to reckon with the impact of the anti-computer strategy on results. In the future it would be necessary to devise methods to describe and research it more closely and how it depends on the strength of engines and depth of search. Problematic is the relative instability of human play, which is clearly illustrated on the graph 20. It however makes coclusions somewhat untrustable. Therefore increase in the number of analyzed moves per player is recommended.

The most difficult part in such analysis works is obviously practical play. There were no satisfactory outcomes regarding that. It was found out that a phenomenon that could be called 'objectivity-practicality bias' is still present in results. Therefore players' results whose difficulty of positions was far from the average must be regarded with caution. The more difficult positions, the bigger the probability that his result according to analysis tourns out to be underrated; and, in case positions far below average, the results will be generally overrated. Previously we saw that practical play is

based on differentials in the three categories: difficulty of positions, type of positions and thinking time. But there is also another interesting, at least a theoretical way. It is based on the fact that moves can be characterized not only by quality, but also by how easily/hard they can be noticed. Some moves are fairly obvious, others can seem quite illogical and at first glance wrong. The fact that there are a lot of positions where the best - often only - move is extremely hard to see, is one of chief reasons why chess is such a difficult and fascinating game. Here we have arrived at the problem of measuring the obviousness of moves. On what basis and how to measure it? We can already now present a simple three-part classification of moves based on obviousness. To the first category belong, for example, moving a piece from the start square, a recapture, giving a check etc. To the second category an exchange sacrifice for a positional advantage, greek gift, a pawn sacrifice for mobility etc. And the third category would include such almost impossible and seemingly absurd moves as 11...Na4!! in Fischer-Byrne 1957; 30. Ba3!! in Botvinnik-Capablanca 1938; 72.Qe5!! in Keres-Fischer 1962; or 24...Nh8!! in Korchnoi-Fischer 1970. However, it is clear that this system is too robust and arbitrary. Assuming we have overcome this limitation, all that would be left is to create a statistical overview of obviousness of moves exceeding a set error threshold. Usually it is easy to distinguish whether a mistake was made due to carelessness or intentionally to create chances. Simple and logical moves hint at negligence, and those that either violate elementary positional principles or give away material is a sign of practical play. It is an interesting prospect worth research. Successful solution of the problem of practical play plays a crucial part in the credibility of conclusions of studies dedicated to chess strength analysis.

The author also compared the 7 most remarklable performances in the chess history. Although Carlsen's expected error was quite far from his official TPR, it nevertheless turned out to be the best of them all. But it is not clear yet. Recently we all witnessed an even more extraordinary performance by Caruana in Sinquefield Cup 2014. Dominguez-Perez in Thessalonikis 2013 and Kramnik in London Classic 2011 were also remarkable performances. There exists a small possibility that their actual quality of play surpasses that of Carlsen. These three performances definitely deserve further scrutiny. The closer a score is to 100% and the fewer games played, the more unreliable TPR value will be. Thus it would be interesting to look into discrepancies between TPR and quality of play as a function of score and the number of games. It is worth to pay attention to tournaments of Lasker and Capablanca in New York in 1924 and 1927. Despite the fact that they both had almost equal chessmetrics TPRs, as shown on the graph 19, that the difference in the quality of play, caused by the objectivity-practicality bias is 205 points. Taking into account the large amount of games and the relatively small timespan between them, it is quite probable that the quality of play of both players closely corresponds to the TPRs. For this reason comparing games of Lasker and Capablanca from the New York tournaments would presumably be a good indicator of the trustability of methods used for analysis.

Results also showed that FIDE rating since 1970 has been inflating with respect to absolute strength, with the average rate about 5 points per decade. It is still relatively modest which may explain why K. W. Regan's work - concluding that there has been no inflation - has not detected this. Regan has analyzed an impressive number of games, but unfortunately the methodology he used was not fully appropriate. The biggest shortcomings were a complete disregard for difficulty of positions and a low search depth due to large amount of data.

There is another use for recording the difficulty of positions besides finding out average excpected error. As previously mentioned, players have an individual tolerance for various aspects of difficulty of positions. Knowledge of which factors of difficulty create more problems and which ones less problems for a certain player is actually very useful. Let's assume a player has determined, via analyzing many games, that he himself is relatively sensitive to the difference between the two best moves; whereas his future opponent has the lowest tolerance for the complexity of positions. He now has to make use of that and to ensure that he only gets such positions on the board where the difference would be as small as possible for him, but as compicated as possible for the opponent. Perhaps the simplest way is to find out the relationship of the degree of different aspects of difficulty against various opening variations. If it should become apparent, for example, that the Sicilian defense has the biggest probability of the occurrence of positions that suit to the player and, at the same time, being relatively less suitable to the opponent, then the reader probably has no difficulty in understanding that, with that, the player gets a serious advantage over his adversary. This is not limited to openings, it is also possible to analyze various types of middle games and endgames to compile an overview of different factors of difficulty by types of positions. Players could use that data to make choices based on tolerances of themselves and their future opponents. The author feels the idea is at least worth paying attention to.