# Using Heuristic-Search Based Engines for Estimating Human Skill at Chess

*Matej Guid*[1] *Ivan Bratko*[1]

Ljubljana, Slovenia

## ABSTRACT

Establishing heuristic-search based chess programs as appropriate tools for estimating human skill levels at chess may seem impossible due to the following issues: the programs' evaluations and decisions tend to change with the depth of search and with the program used. In this research, we provide an analysis of the differences between heuristic-search based programs in estimating chess skill. We used four different chess programs to perform analyses of large data sets of recorded human decisions, and obtained very similar rankings of skill-based performances of selected chess players using any of these programs at various levels of search. A conclusion is that, given two chess players, all the programs unanimously rank one player to be clearly stronger than the other, or all the programs assess their strengths to be similar. We also repeated our earlier analysis with the program CRAFTY of World Chess Champions with currently one of the strongest chess programs, RYBKA 3[2], and obtained qualitatively very similar results as with CRAFTY. This speaks in favour of computer heuristic search being adequate for estimating skill levels of chess players, despite the above stated issues.

## 1. INTRODUCTION

In sports and games, rating systems of various kinds are a widely accepted method for estimating skill levels of the players. These rating systems are based on outcomes of direct competitions only. Our approach is different: we assess skill level at chess by applying chess engines to analyse particular positions and moves played.

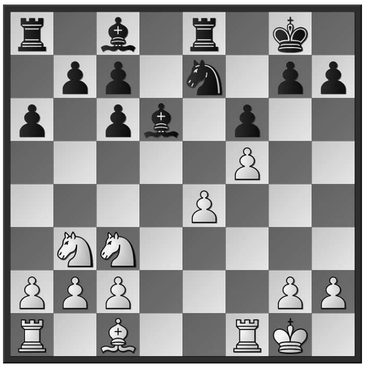
In Guid and Bratko (2006), we carried out a computer analysis of individual moves in World Chess Championship matches with the chess program CRAFTY in an attempt to assess as objectively as possible one aspect of the playing strength of chess players of different eras. Also, a method was designed for assessing the complexity of a position, in order to take into account the different playing styles of the players and the difficulty of the positions they are faced with. Subsequent statistical analysis (Guid, Perez, and Bratko, 2008) demonstrated that, at least for pairs of the players whose scores differed significantly, it is not very likely that the relative rankings of the champions according to the criterion considered would change if a stronger chess program was used for the analysis. In current article, we verify this claim empirically by applying three chess programs stronger than CRAFTY for the same type of computer analysis of the players' skills.

Establishing heuristic-search based computer programs as an appropriate tool for estimating performance of chess players in executing their tasks during a chess game may seem impossible, since it is well known that both programs' evaluations and programs' decisions tend to change as the depth of search increases. It is very likely that in the future chess programs will continue to change their decisions by searching more deeply, and that the frequencies of changes will remain significant for all feasible search depths.[3] Not to mention that the cease of decision changes at some particular search depth does not guarantee optimal play at all. Also, it is even unclear what "optimal" is? For a human, it is not necessarily the shortest path to win, but a kind of "easiest" and most reliable one, that is one that minimizes the risk – the probability for a human to make a mistake. It is, however, difficult to determine such probabilities. Moreover, different programs typically assign different evaluations to

[1] Artificial Intelligence Laboratory, Faculty of Computer and Information Science, University of Ljubljana.

[2] [We refer to other articles on RYBKA in this issue, and accept the similarity. -Ed.]

[3] For the typical chess programs' rates of changed decisions with depth of search refer to (Steenhuisen, 2005) and (Guid and Bratko, 2007).



| Program | Evaluation |
|---|---|
| CHESSMASTER 10 | 0.15 |
| CRAFTY 19.19 | 0.20 |
| CRAFTY 20.14 | 0.08 |
| DEEP SHREDDER 10 | -0.35 |
| DEEP SHREDDER 11 | 0.00 |
| FRITZ 6 | -0.19 |
| FRITZ 11 | 0.07 |
| RYBKA 2.2n2 | -0.01 |
| RYBKA 3 | -0.26 |
| ZAPPA 1.1 | 0.13 |

**Figure 1**: Lasker-Capablanca, St. Petersburg 1914, position after White's $12^{th}$ move. The table on the right shows backed-up heuristic evaluations obtained by various chess programs, when evaluating the diagrammed chess position using 12-ply search.

a given position, even when using the same depth of search (see Fig. 1). Which program is therefore to be the most trusted one as the estimator: the strongest one in terms of the programs' competition strength, or perhaps the one that is most equally unfair to all the players that are a subject of evaluation? These issues seem like an obstacle on the way to establishing heuristic-search based methods as competent problem-solving performance estimators, especially in complex games such as chess.

Our approach to estimating skill levels is therefore aimed at assessing the quality of play regardless of the game score, and assumes using fallible estimators (as opposed to infallible estimator such as chess tablebases (Thompson, 1986)). A related approach was introduced by Haworth (2007): he defined a mapping of the apparent player's skill into a Referent Agent Space, using a stochastic agent and a Bayesian inference method. The results of the experimental analysis in using chess programs TOGA and SHREDDER at search depth of 10 plies (Di Fatta, Haworth, and Regan, 2009; Haworth, Regan, and Di Fatta, 2010) showed that the inferred probability distributions of the apparent skills are able to discriminate players' performances in different Elo ranges. Although both approaches require further work on how to take into account the differences between players in the average difficulty of the positions encountered in their games, the experiments carried out nevertheless demonstrated viability of estimating skills based on computer heuristic evaluation.

In this article, we are particularly interested in analysing the behaviour of different chess programs at different levels of search, when the same type of computer analysis of chess-players' skills as in Guid and Bratko (2006) and Guid *et al.* (2008) is conducted. The remainder of the article is organized as follows. In Section 2, we conduct experiments where three chess champions whose skill-based performances were previously established to deviate significantly from the others are selected for the analysis by chess programs of different competition strengths, and compared to a control group of more than 20,000 randomly picked positions of other chess champions. Section 3 presents the results of the experiments. In Section 4, we analyse an impact of the *monotonicity property* of heuristic evaluation functions on the results of the computer analysis. In Section 5, we present results of the computer analysis of World Chess Champions with currently one of the strongest chess programs, RYBKA 3. A brief discussion about establishing computer heuristic search as an appropriate tool for estimating chess-players' skills follows. We summarise our results in Section 7.

## 2. EXPERIMENTAL DESIGN

Both of the aforementioned approaches to estimating skill levels in complex games have something in common, besides that they are based on computer analysis of player's actions: they both assume (implicitly or explicitly) that average differences in computer evaluations between the player's moves and the computer's moves provide a solid ground for sensible benchmarking of the players, when a sufficiently large amount of data for the analysis is available. That is, regardless how this measure is further refined to take into account the full context of the player's decisions, these average differences alone are expected to have good chances to produce rather sensible rankings of the players, even when using different programs as estimators of the players' skills.

Guid *et al.* (2008) were particularly interested in observing to what extent the World Chess Champions' rankings, based on these differences, are preserved at different search depths. The computer analysis was based on the evaluation of the games played by the chess champions in the classical championship matches between 1886 (Steinitz-Zukertort) and 2006 (Kramnik-Topalov). The aim was to assess how reliable CRAFTY (or, by extrapolation, any other chess program) is as a tool for the comparison of chess players. It was shown that at least for the players whose scores differ significantly, the rankings are practically the same at different levels of search. This finding suggested that using a program stronger than CRAFTY would lead to similar rankings of the players.

Also, some champions whose rankings significantly deviate from the others were identified: Capablanca, Euwe, and Steinitz. Their relative rankings among all the champions were preserved at each level of search. The stability of their rankings was further supported by the statistical analysis of the results and from the fact that Capablanca had the best score in 95% of all the subset-depth combinations in 100 samples consisting of 500 randomly chosen positions (Guid *et al.*, 2008).

To study whether different heuristic-search based programs can be used as estimators of skill levels, we will now focus on these three players and check whether other chess programs rank them in the same order as CRAFTY did. We will observe variations with depth of average differences between player's and program's decisions on a large subset of randomly chosen positions from the World Chess Championship matches using three chess programs stronger than CRAFTY, namely: SHREDDER, RYBKA 2, and RYBKA 3.

In order to provide a relevant comparison of the strength of the four programs, relative to each other, we give in Table 1 the publicly available information from the Swedish (SSDF) Rating List (2009-04-10) (Karlsson, 2008), where ratings of several state-of-the-art chess programs are published. The ratings on this list are obtained after many games are played on the tournament level (40 moves in 2 hours followed by 20 moves in each following hour) between the programs, supervised by members of the Swedish Chess Computer Association (SSDF). 'Games' stands for the number of games on which the rating is based, and 'Against' for the average rating of opponents. The '+' and '-' indicate the 95% confidence intervals. Although the versions of the programs slightly differ from the ones that we used in our experiments,[4] the listed ratings should give sufficient information about their strength, in particular relative to each other.

**Table 1**: Comparison of the four programs that we used in our experiments according to the SSDF rating list.

| Program | Rating | + | - | Games | Win % | Against |
|---|---|---|---|---|---|---|
| DEEP RYBKA 3 | 3073 | 44 | -44 | 253 | 55 | 3039 |
| RYBKA 2.3.1 Arena | 2923 | 23 | -23 | 920 | 53 | 2904 |
| SHREDDER 10 UCI | 2827 | 20 | -20 | 1246 | 58 | 2769 |
| CRAFTY 19.17 | 2527 | 41 | -44 | 304 | 30 | 2677 |

All positions for the analysis were taken from World Chess Championship matches. Besides analysing the moves of the three players, we also randomly picked more than 20,000 positions of the rest of the players as a control group. We used the same methodology for determining the rankings of World Chess Champions as in Guid *et al.* (2008). That is, searches to various depths were conducted (including quiescence search to obtain stable evaluations). Then the *scores* obtained by the program are the average differences between computer evaluations of the players' choices of moves and the computer's choices at each particular search depth. Based on the players' scores, the rankings of the players are obtained so that a lower score results in a better ranking. As we advocated in Guid *et al.* (2008), these scores that are relative to the computer used, have good chances to produce sensible rankings of the players.

Moves where both the move played and the move suggested by the computer had an evaluation outside the interval [-2, 2], were discarded and not taken into account in the calculations of the scores of the players (this was done at each depth separately), due to the reasons given in Guid and Bratko (2006).[5] Since the thresholds of this interval were set arbitrarily, the following question arises. What is the impact of the size of this interval on the rankings? In order to answer this question, we varied the size of this interval in the range from [-1, 1] to [-3, 3] by the step of 0.5 in both directions. Chess program CRAFTY was used in this experiment. The four curves so obtained for each of the players are given in Figure 2. They demonstrate that the size of this interval does *not*

[4] CRAFTY 19.2, RYBKA 2.2n2 32-bit, RYBKA 3 1-cpu 32 bit, and DEEP SHREDDER 10 UCI were used in our experiments.

[5] In clearly won positions players are tempted to play a simple safe move instead of a stronger, but risky one. Such moves are, from a practical viewpoint, justified. Taking them into account would wrongly penalize players that used this legitimate approach trying (and usually succeeding) to obtain desired result.

have a significant impact on the relative rankings of the players in question. None of the four different curves for each particular player overlaps with any curve of another player or the control group of other players.

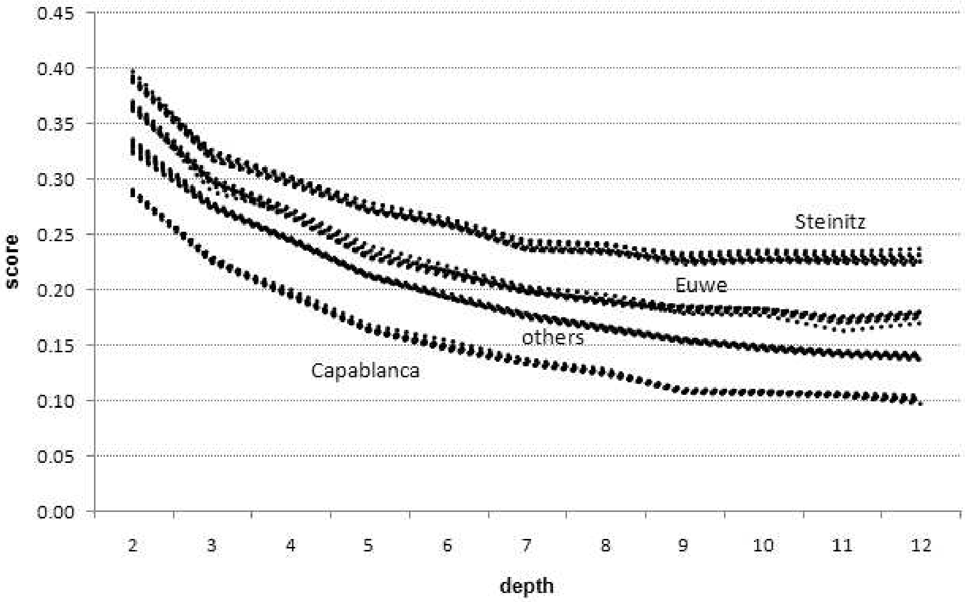


**Figure 2**: The scores of Capablanca, Euwe, Steinitz, and other players, obtained on a large subset of games from World Chess Champion matches, using CRAFTY. For each player and search depth, there are several data points that correspond to different lost/won thresholds. Clearly, changing these thresholds that define when a position was in the calculations of the scores due to the reasons given in Guid and Bratko (2006), does not affect the rankings of the players.

## 3. EXPERIMENTAL RESULTS

The results are given in Figures 3, 4, and 5, for SHREDDER, RYBKA 2, and RYBKA 3, respectively. The relative ranking of Capablanca, Euwe, Steinitz, and the players in the control group are preserved at all depths using any of the programs at any level of search. These experimental findings reject possible speculations that using chess programs stronger than CRAFTY for the analysis could lead to completely different results, and that Capablanca's result is merely a consequence of his style being similar to CRAFTY's style. We also view these results as another confirmation that in order to obtain a sensible ranking of the players, it is not necessary to use a computer that is stronger than the players themselves.

The scores of the players tend to decrease with increasing search depth regardless of the program used. This is probably the consequence of the computers' evaluations becoming more reliable with an increasing depth of search, and therefore on average achieving a better match with the players. In some cases, however, at greater search depths the scores' tendencies are reversed, so that they start to increase with the depth of search. This phenomenon is discussed in Section 4.

The experimental results presented in Figures 3, 4, and 5 not only confirm that the scores are not invariable for the same program at different depths of search, the scores also differ significantly when using different programs. This is most clearly seen in Fig. 6, where average scores of all the players, obtained on the same large subset of games from World Chess Champion matches, are given for the three programs, and compared to average scores of all the players according to CRAFTY. While the scores of SHREDDER are very similar to the scores of CRAFTY, the scores of the two RYBKAs differ considerably from CRAFTY and SHREDDER.

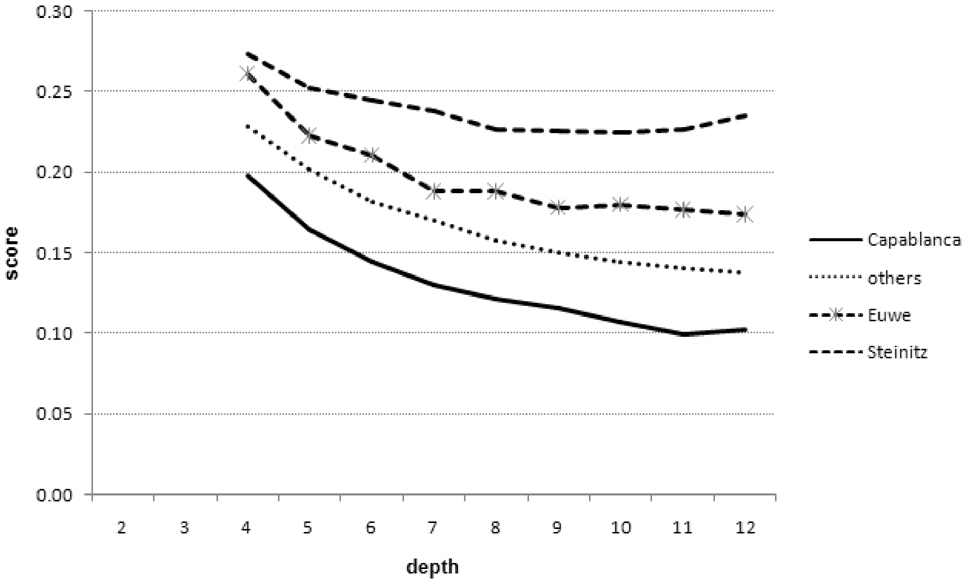


**Figure 3**: The scores of Capablanca, Euwe, Steinitz, and other players, obtained on a large subset of games from World Chess Champion matches, using SHREDDER.

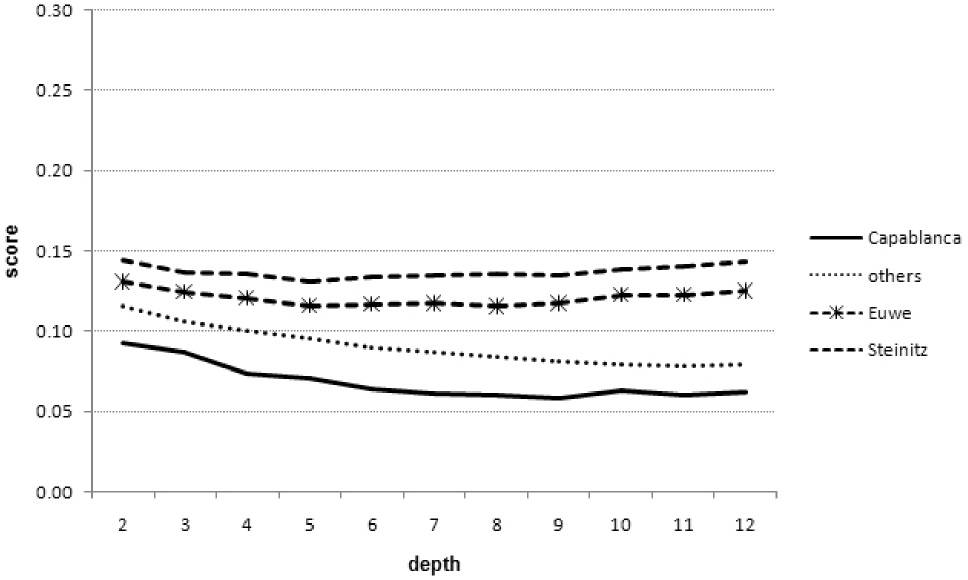


**Figure 4**: The scores of Capablanca, Euwe, Steinitz, and other players, obtained on a large subset of games from World Chess Champion matches, using RYBKA 2.

## 4. SCORES AND MONOTONICITY PROPERTY OF HEURISTIC EVALUATION FUNCTIONS

The shape of the curves that represent the scores of Steinitz in Figures 3 to 5 is particularly interesting. The previously observed decreasing tendency of the scores with increasing search depth (Guid *et al.*, 2008) does *not*

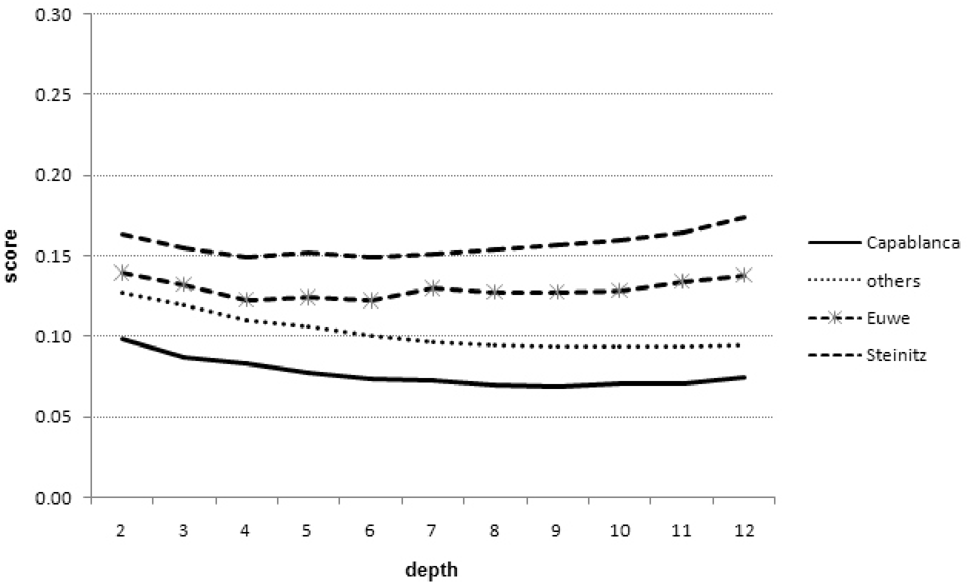


**Figure 5**: The scores of Capablanca, Euwe, Steinitz, and other players, obtained on a large subset of games from World Chess Champion matches, using RYBKA 3.

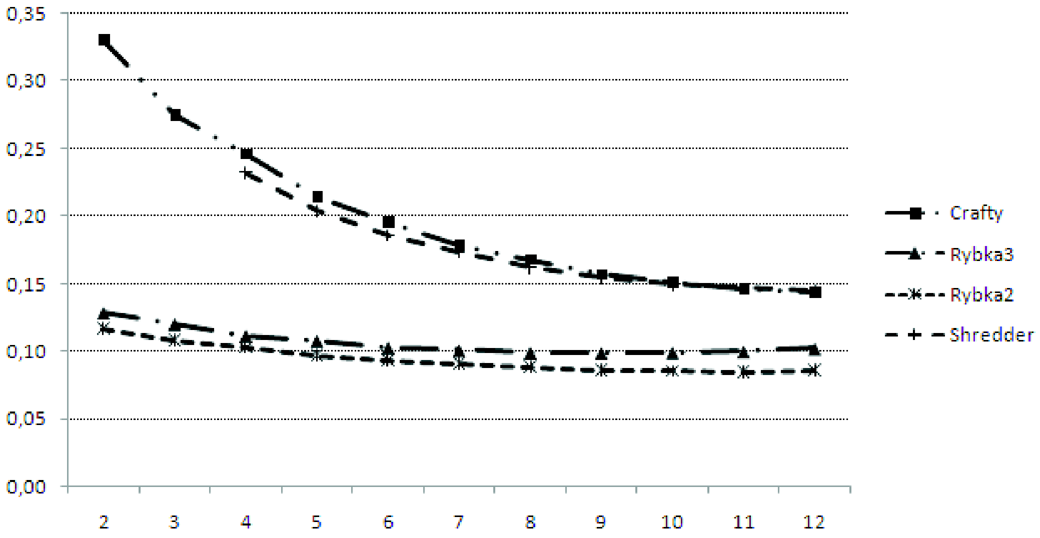


**Figure 6**: Comparison of average scores of all the players (including the positions of the control group), obtained by CRAFTY, SHREDDER, RYBKA 2, and RYBKA 3.

always hold here – there is even an *increasing* tendency of the scores of Steinitz in Figures 3, 4, and 5 when the search approaches depth 12. The same phenomenon can be observed in the average scores of all the players using the strongest of the four programs, RYBKA 3 (see Figure 6): while the scores slightly decrease at lower search depths, they tend to increase at higher search depths.

We believe that this phenomenon is a consequence of the *monotonicity property* of successful heuristic evaluation functions for games (Guid, 2010). Namely, that backed-up values of the nodes in the search space change monotonically with the depth of search. Guid (2010) demonstrated that in positions with a winning advantage the backed-up evaluations of chess programs on average monotonically increase with increasing search depth, and that the backed-up evaluations of better moves (*i.e.*, higher-valued moves according to the program) on

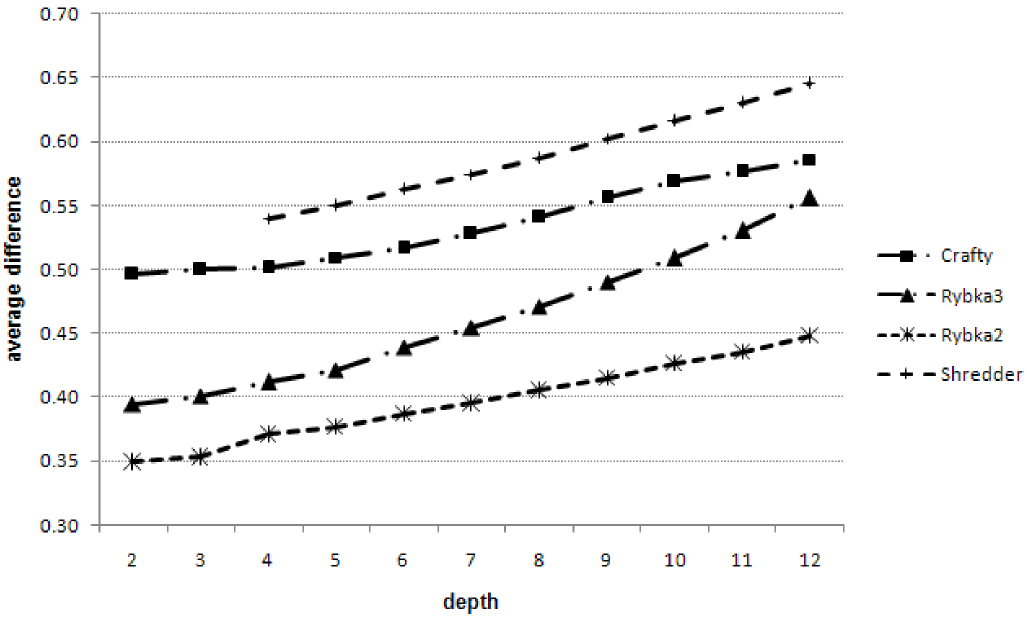


**Figure 7**: Comparison of average differences between best two moves chosen by the corresponding program in all positions from World Chess Championship, obtained by CRAFTY, SHREDDER, RYBKA 2, and RYBKA 3.

average increase more rapidly in such positions.[6] Therefore, in positions with a decisive advantage the differences between the backed-up evaluations of candidates for the best move according to the program are likely to become bigger with increasing search depth. This can be clearly seen in Figure 7, where the average differences between the best two moves in positions according to the particular program are shown at each depth of search: for each program they increase monotonically with search depth. All positions from World Chess Championship matches were taken into account in this experiment.

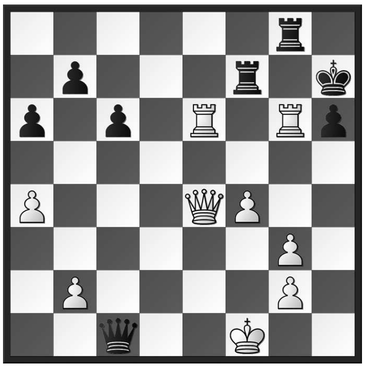
The cost of not choosing the best move as seen by the program therefore on average increases with increasing search depth. This finding suggests the following explanation for the increasing tendency of the scores with depth in some cases. The *decreasing* tendency of the scores at the shallower search depths are the consequence of increasingly better program choices (speaking on average) with depth of search. However, at higher search depths the higher costs of not choosing the best move according to the program leads to the *increased* scores due to the monotonicity property of heuristic evaluation functions.

Figure 7 also demonstrates that the absolute value of the average differences between best two moves varies with respect to the program used. This finding provides an explanation why the scores of the four programs at particular search depths are not on the same scale. That is, as Figure 6 and Figure 7 clearly suggest, the meaning of particular score values vary with respect to the program that is used for evaluating a player's skills. Moreover, the magnitude of the scores obtained by the four programs is not correlated with the programs' competition strengths (compare Figure 6 and Figure 7 to Table 1).

## 5. COMPUTER ANALYSIS OF WORLD CHESS CHAMPIONS REVISITED WITH RYBKA 3

In this section, we present the results of computer analysis of World Chess Champions obtained by RYBKA 3. We used the same methodology and analysed the same matches as in Guid *et al.* (2008), in order to make the comparison of the results obtained by chess programs CRAFTY and RYBKA 3 as relevant as possible. That is, games for the title of World Chess Champion, in which the fourteen classic World Champions (from Steinitz to Kramnik) contended for or were defending the title, were selected for the analysis. Each position occurring in

[6] More precisely: with increasing search depth, backed-up evaluations of won positions (in theoretical sense: White wins providing optimal play by both sides) will on average be increasing, evaluations of lost positions will be decreasing, while evaluations of positions with game-theoretical value draw will be converging towards 0 and search will eventually end in terminal nodes that represent theoretical draw.



| Program | Kf1-e2 | Kf1-f2 |
|---|---|---|
| CHESSMASTER 10 | 3.63 | 0.00 |
| CRAFTY 19.19 | 2.33 | 0.00 |
| CRAFTY 20.14 | 2.34 | 0.00 |
| DEEP SHREDDER 10 | 2.44 | 0.00 |
| DEEP SHREDDER 11 | 2.41 | 0.00 |
| FRITZ 6 | 1.00 | 0.00 |
| FRITZ 11 | 1.74 | 0.00 |
| RYBKA 2.2n2 | 1.65 | 0.00 |
| RYBKA 3 | 5.12 | 0.00 |
| ZAPPA 1.1 | 3.33 | 0.00 |

**Figure 8**: Capablanca-Alekhine, World Chess Championship match (game 27, position after Black's 37th move), Buenos Aires 1927. White is winning after 38. Kf1-e2!, for example: 38. ... Qc1xb2+ 39. Ke2-f3 Qb2-b3+ (or 39. ... Qb2-c3+ 40. Kf3-g4!) 40. Kf3-f2! and the white King safely returns to h2, avoiding any further checks. However, Capablanca played 38. Kf1-f2?? and after 38. ... Qc1-d2+! had to agree to a draw, since there is no escape from perpetual check now. *How much should White be "punished" for his mistake?* The table on the right shows backed-up heuristic evaluations obtained by various chess programs, when evaluating the moves 38. Kf1-e2 (computer's choice) and 57. Kf1-f2 (move played) using 12-ply search. While all engines agree that the position is equal after 38. Kf1-f2, there are vast differences among the evaluations of 38. Kf1-e2.

these games after move 12 was iteratively searched to depths ranging from 2 to 10 ply.[7] Again, based on the players' scores the rankings of the players were obtained so that a lower score results in a better ranking.

There was however one, but nonetheless important modification to the previously adopted methodology. We observed that in some (albeit rare) cases a player may get too severely "punished" for his (or her) mistake. This often has nothing to do with the magnitude of a mistake as seen through the eyes of, say, a human grandmaster. Let us demonstrate this by the example given in Figure 8. The white player made a big mistake on his 38th move that changed the assessment of the position from a winning into a drawn one. Of course, there were many such turnovers in World Chess Championship matches. However, the penalty for such a mistake completely depends on the properties of a program's evaluation function, which may in some cases assign unreasonably high (from a human's perspective) numerical values to the difference between a computer's choice and the move played. Obviously, the computers lack consistency in such cases: on some occasions a mistake by the player may be much smaller from the viewpoint of a human expert, but receives an incomparably higher "penalty," and vice versa. The monotonicity property of heuristic evaluation functions further magnifies this effect, and such inconsistencies may unfairly affect the overall rankings of the players.

To reduce the negative impacts of these inconsistencies in the programs' evaluations, we determined a maximal error (as seen by the engine) for a single move. This value was arbitrarily set to 3.00 (*i.e.*, 300 *centipawns*). Through the eyes of a human expert this number represents a huge mistake.[8] Therefore, if the difference between the computer's choice and the move played in a particular position was higher than 3.00, the difference was automatically set to 3.00. Although it is difficult to determine the most suitable value, we believe that this modification, using any reasonable value for a human expert's estimation of a huge mistake, represents an improvement of the methodology presented in Guid and Bratko (2006) and Guid *et al.* (2008).

The results of the computer analysis by RYBKA 3 are presented in Figure 9. They closely resemble the ones obtained by CRAFTY. Again it holds for almost all depths: rank(Capablanca) < rank(Kramnik) < rank (Karpov, Kasparov) < rank(Petrosian) < rank(Botvinnik) < rank(Euwe) < rank(Steinitz). However, there are nevertheless some obvious differences between the results obtained by the two programs. Firstly, CRAFTY ranked Karpov better than Kasparov at all depths: just the opposite holds with RYBKA 3. Nevertheless, the scores of both (and also many other) players are rather similar to each other, which is not surprising provided that it is likely that many World Chess Champions had a rather similar chess strength. Secondly, the performance of Fischer as seen

[7] Searches to 11 and 12 plies were omitted in order to shorten the computational time required for such an extensive analysis.

[8] Although this modification makes it harder to compare directly the newly obtained results to the ones presented in Guid *et al.* (2008), not applying it would not resolve the issue either: the inconsistencies in assigning high numerical values to particular positions by two different programs are highly unlikely on the same scale. Applying the same modification to CRAFTY's analysis of World Chess Champions would not make it any easier to compare the results of the two programs, since CRAFTY's 300 centipawns are not directly comparable to RYBKA's 300 centipawns at the same level of search. An interested reader can find more details on these topics in Guid (2010, Chapter 8).

by RYBKA 3 is significantly better in comparison with the results obtained by CRAFTY. This finding is associated with the following (so far unanswered) question addressed in Guid *et al.* (2008): Does a program's style of play exhibit preference for the styles of any particular players? [9] The comparison of the results of computer analysis of the champions obtained by CRAFTY and RYBKA 3 suggests that a program's style of play may affect the rankings (as in the aforementioned cases), but only to a limited extent.

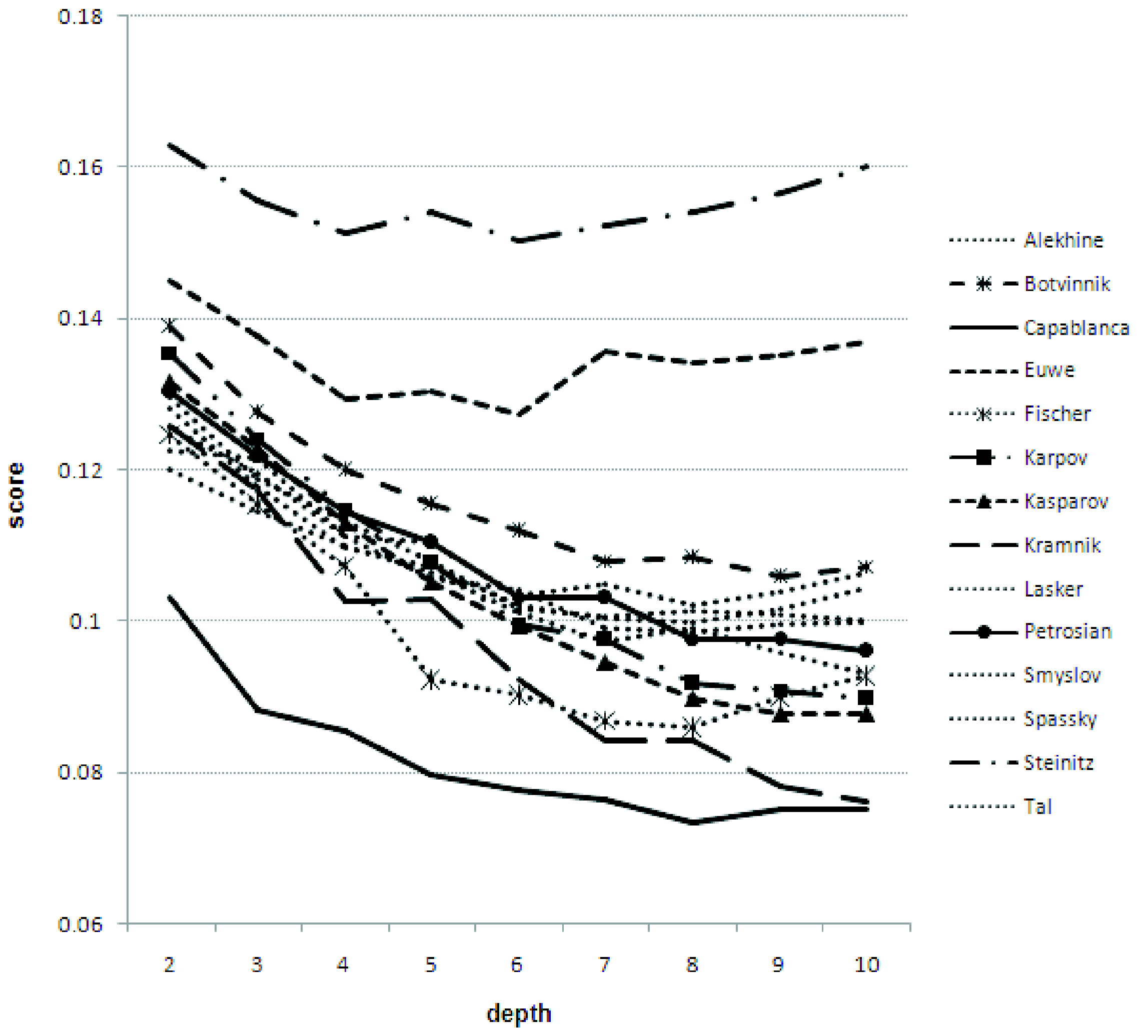


**Figure 9**: The scores of World Chess Champions obtained by chess program RYBKA 3 at various depths of search. Based on the players' scores the rankings of the players were obtained. For almost all depths it holds that rank(Capablanca) < rank(Kramnik) < rank (Kasparov) < rank (Karpov) < rank(Petrosian) < rank(Botvinnik) < rank(Euwe) < rank(Steinitz).

[9] For example, the evaluation functions of two programs may value particular features of chess positions (*e.g.*, the presence of opposite coloured bishops) differently, in a more similar or in a less similar way as a particular human player.

## 6. DISCUSSION

Rating systems of various kinds are a widely accepted method for estimating skill levels in sports and games, but they are based on outcomes of competitive activities only and may not reflect the true skill level that a player demonstrates during these activities. In chess, for example, a complete beginner and a master may both score equally against a grand master. Moreover, a single mistake can ruin a well-played game, and the same game result between two chess players can be obtained by a completely different quality of play. In general, a statistical analysis of results in competition merely reflects a player's success in competition, but not directly the skills that the player demonstrates.

An alternative to the outcome-based ratings is the estimation of skill levels based on an analysis of the actions performed by the player. A trustworthy estimator of demonstrated skills is required for this purpose. There are many arguments in favour of computer programs as such estimators. In contrast to humans, they: (1) have an enormous computing power, (2) use numerical values as evaluations, (3) adhere to the same rules all the time, and (4) are not influenced by emotions. Computer programs therefore have a capability of being more consistent than human observers, and can deal with incomparably more observations in a limited time.

State-of-the-art chess programs are based on heuristic search. Their strength increased drastically in the past few decades; they are nowadays superior even to the strongest human grandmasters - already surpassing them in many aspects (Kasparov, 2003-2006). It is therefore desirable to establish computer heuristic search as an appropriate tool for estimating chess-players' skills (or chess strength). This would provide additional insights into chess players, *e.g.*, how they perform in different types of game positions or in different phases of the game, and make a comparison of different players that never have (or had) a chance to compete against each other at the chessboard possible.

One of the important (and still unresolved) issues is: How to take into account the differences between players in the average difficulty of the positions encountered in their games? In Guid and Bratko (2006), we devised a heuristic-search based method to assess average difficulty of positions used for estimating the champions' performance and presented the results of applying this method in order to compare chess players of different playing styles. Those results suggested that Capablanca's outstanding score in terms of low average differences in computer evaluations between the player's moves and the computer's moves should be interpreted in the light of his playing style that tended towards low complexity positions. However, we believe that further work on the above question is required.

## 7. CONCLUSIONS

We investigated the appropriateness of heuristic-search based programs for estimating skill levels in game playing. This task is seemingly impossible due to the fact that the programs' evaluations and decisions tend to change with search depth and with respect to the program used. We were particularly interested in analysing behaviour of different chess programs at different levels of search, when the same type of computer analysis of chess-players' skills as in Guid and Bratko (2006) and Guid *et al.* (2008) is conducted. That is, chess players are benchmarked against each other based on their scores, calculated as average differences in the evaluation values between the moves actually played and the best moves found by the program.

The results presented in this paper demonstrated that heuristic-search based chess programs can nonetheless act as reliable estimators of skill levels, despite of the above mentioned issues. We selected a large data set of recorded decisions of three chess players whose skill levels were previously established to significantly deviate from one another (supplemented by a control group of 20,000 recorded decisions of other players) and compared their scores at different levels of search using four chess programs of different competition strengths. The relative rankings of the three players and the players in the control group remained preserved at all depths using any of the four programs at any level of search.

We also demonstrated that the scores may differ significantly when using different programs, and provided the following explanation regarding the magnitude of the scores with respect to the search depth. The *decreasing* tendency of the scores with an increasing search depth (typical at the shallower search depths) is a consequence of increasingly better choices by the program. At higher search depths, however, the higher cost of *not* choosing the best move according to the program may even lead to *increased* scores due to the *monotonicity property* of heuristic evaluation functions (Guid, 2010).

Subsequently, the computer analysis of the World Chess Champions was repeated using one of the strongest chess programs, RYBKA 3. The results of this analysis are qualitatively very similar to the ones obtained by the chess program CRAFTY. This experiment was closely related to the following question: does a program's style of play exhibit preference for styles of any particular players? The results suggest that a program's style of play may affect the rankings, but only to a limited extent.

Once again, to avoid possible misinterpretation of the presented work, it should be noted that this article is *not* concerned with the question of how appropriate this particular measure of the playing strength (deviation of player's moves from computer-preferred moves) is as a criterion for comparing chess players' ability in general. It is only one of many sensible criteria of this kind. For a more accurate estimation of the players' skills it would be useful, for example, to also take into account the time spent on particular decisions (this information is usually not recorded for chess games). Moreover, it would be desirable to take into account the cognitive difficulty of chess positions that the players were facing. Nevertheless, the relatively good preservation of the rankings at different levels of search using various programs of different competition strengths represent a solid ground on our way to discover increasingly better methods for establishing heuristic-search based computer programs as reliable estimators of skill levels in game playing.

## 8. REFERENCES

Di Fatta, G., Haworth, G., and Regan, K. (2009). Skill rating by Bayesian inference. *CIDM*, pp. 89–94.

Guid, M. (2010). *Search and Knowledge for Human and Machine Problem Solving.* Ph.D. thesis, University of Ljubljana, Slovenia.

Guid, M. and Bratko, I. (2006). Computer Analysis of World Chess Champions. *ICGA Journal*, Vol. 29, No. 2, pp. 3–14.

Guid, M. and Bratko, I. (2007). Factors Affecting Diminishing Returns for Searching Deeper. *ICGA Journal*, Vol. 30, No. 2, pp. 65–73.

Guid, M., Perez, A., and Bratko, I. (2008). How Trustworthy is CRAFTY's Analysis of World Chess Champions? *ICGA Journal*, Vol. 31, No. 3, pp. 131–144.

Haworth, G. (2007). Gentlemen, stop your engines! *ICGA Journal*, Vol. 30, No. 3, pp. 150–156.

Haworth, G., Regan, K., and Di Fatta, G. (2010). Performance and Prediction: Bayesian Modelling of Fallible Choice in Chess. *Advances in Computers Games*, Vol. 6048 of *Lecture Notes in Computer Science*, pp. 99–110, Springer.

Karlsson, T. (2008). The Swedish Rating List. *ICGA Journal*, Vol. 31, No. 4, p. 255.

Kasparov, G. (2003-2006). *My Great Predecessors, Parts 1-5.* Everyman Chess, London.

Steenhuisen, J. R. (2005). New Results in Deep-Search Behaviour. *ICGA Journal*, Vol. 28, No. 4, pp. 203–213.

Thompson, K. (1986). Retrograde Analysis of Certain Endgames. *ICCA Journal*, Vol. 9, No. 3, pp. 131–139.