How Trustworthy is CRAFTY'S Analysis of Chess Champions?

Matej Guid[1], Aritz Pérez[2], and Ivan Bratko[1]

[1]Univ. of Ljubljana, Slovenia, and [2]Univ. of San Sebastián, Spain

Abstract. Guid and Bratko carried out a computer analysis of games played by World Chess Champions as an attempt at an objective assessment of chess playing strength of chess players of different times. Chess program CRAFTY was used in the analysis. Given that CRAFTY's official chess rating is lower than the rating of many of the players analysed, the question arises to what degree that analysis could be trusted. In this paper we investigate this question and other aspects of the trustworthiness of those results. Our study shows that it is not very likely that the ranking of at least the two highest-ranked players would change if (1) a stronger chess program was used, or (2) if the program would search deeper, or (3) larger sets of positions were available for the analysis.

1 Introduction

The emergence of high-quality chess programs provided an opportunity of a more objective comparison between chess players of different eras who never had a chance to meet across the board. Recently Guid and Bratko [4] published an extensive computer analysis of World Chess Champions, aiming at such a comparison. It was based on the evaluation of the games played by the World Chess Champions in their championship matches. The idea was to determine the chess players' *quality of play* (regardless of the game score), which was evaluated with the help of computer analyses of individual *moves* made by each player. The winner according to the main criterion, where average deviations between evaluations of played moves and best-evaluated moves according to the computer were measured, was Jose Raul Capablanca, the 3rd World Champion, which to many came as a surprise (although Capablanca is widely accepted as an extremely talented and a very accurate player).

A version of that article was republished by a popular chess website, ChessBase.com [3], and various discussions took place at different blogs and forums across the internet, while the same website soon published some interesting responses by various readers from all over the world, including some by scientists [2]. A frequent comment by the readers could be summarised as: "A very interesting study,

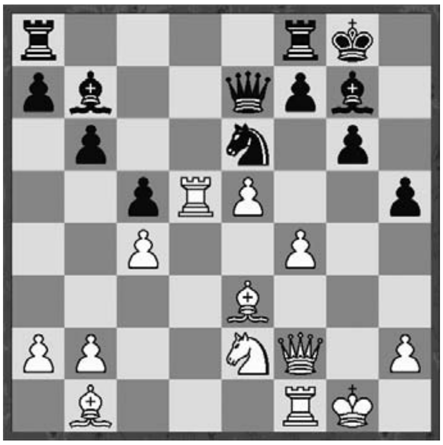
[1] Artificial Intelligence Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Slovenia. Email: {matej.guid,bratko}@fri.uni-lj.si.

[2] Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country. Email: aritz@si.ehu.es.

but it has a flaw in that program Crafty, whose rating is only about 2620, was used to analyse the performance of players stronger than this. For this reason the results cannot be useful". Some readers speculated further that the program will give better ranking to players that have a similar strength to the program itself. In more detail, the reservations by the readers included three main objections to the used methodology:

- the program used for analysis was too weak,
- the depth of the search performed by the program was too shallow[3],
- the number of analysed positions was too low (at least for some players).

In this paper we address these objections in order to determine how reliable Crafty (or any other fallible chess program) is as a tool for comparison of chess players, using the suggested methodology. In particular, we were interested in observing to what extent is the ranking of the players preserved at different depths of search. Our results show, possibly surprisingly (see Fig. 1), that at least for the players whose score differentiate enough from the others (as is the case for Capablanca and Kramnik on one side of the list, and Euwe and Steinitz on the other) the ranking remains preserved, even at very shallow search depths.



depth	best move	evaluation
2	Bxd5	-1.46
3	Bxd5	-1.44
4	Bxd5	-0.75
5	Bxd5	-1.00
6	Bxd5	-0.60
7	Bxd5	-0.76
8	Rad8	-0.26
9	Bxd5	-0.48
10	Rfe8	-0.14
11	Bxd5	-0.35
12	Nc7	-0.07

Fig. 1. Botvinnik-Tal, World Chess Championship match (game 17, position after white's 23rd move), Moscow 1961. In the diagram position, Tal played 23…Nc7 and later won the game. The table on the right shows Crafty's evaluations as results of different depths of search. As it is usual for chess programs, the evaluations vary considerably with depth. Based on this observation, a straightforward intuition suggests us that by searching to different depths, different rankings of the players would have been obtained. However, as we demonstrate in this paper, the intuition may be misguided in this case, and statistical smoothing prevails.

It is well known for a long time that strength of computer chess programs increases with search depth. Already in 1982, Ken Thompson [8] compared programs that searched to different search depths. His results show that searching to only one ply

[3] Search depth in the original study was limited to 12 plies (13 plies in the endgame) plus quiescence search.

deeper results in more than 200 rating points stronger performance of the program. Although later it was found that the gains in the strength diminish with additional search, they are nevertheless significant at search depths up to 20 plies [6]. The preservation of the rankings at different search depths would therefore suggest not only that the same rankings would have been obtained by searching deeper, but also that using a stronger chess program would not affect the results significantly, since the expected strength of CRAFTY at higher depths (e.g. at about 20 plies) are already comparable with the strength of the strongest chess programs, under ordinary tournament conditions at which their ratings are measured (see [7] for details).

We also studied how the scores and the rankings of the players would deviate if smaller subsets of positions were used for the analysis, and whether the number of positions available from world championship matches suffices for successful ranking of the World Champions.

2 Method

We used the same methodology as Guid and Bratko [4] did in their study. Games for the title of "World Chess Champion", where the fourteen classic World Champions contended for or were defending the title, were selected for analysis. Each position occurring in these games after move 12 was iteratively searched to depths 2 to 12 ply. Search to depth *d* here means *d* ply search extended with quiescence search to ensure stable static evaluations. The program recorded best-evaluated moves and their backed-up evaluations for each search depth from 2 to 12 plies (Fig. 2). As in the original study, moves where both the move made 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. In such clearly won positions players are tempted to play a simple safe move instead of a stronger, but risky one. The only difference between this and the original study regarding the methodology, is in that the search was not extended to 13 plies in the endgame. Obviously the extended search was not necessary for the aim of our analysis: to obtain rankings of the players at the different depths of search.

The average differences between evaluations of moves that were played by the players and evaluations of best moves suggested by the computer were calculated for each player at each depth of the search. The results are presented in Fig. 3.

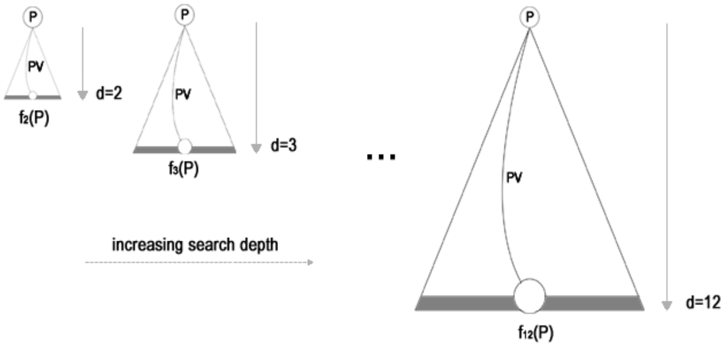


Fig. 2. In each position, we performed searches to depths from 2 to 12 plies extended with quiescence search to ensure stable static evaluations. Backed-up evaluations of each of these searches were used for analysis.

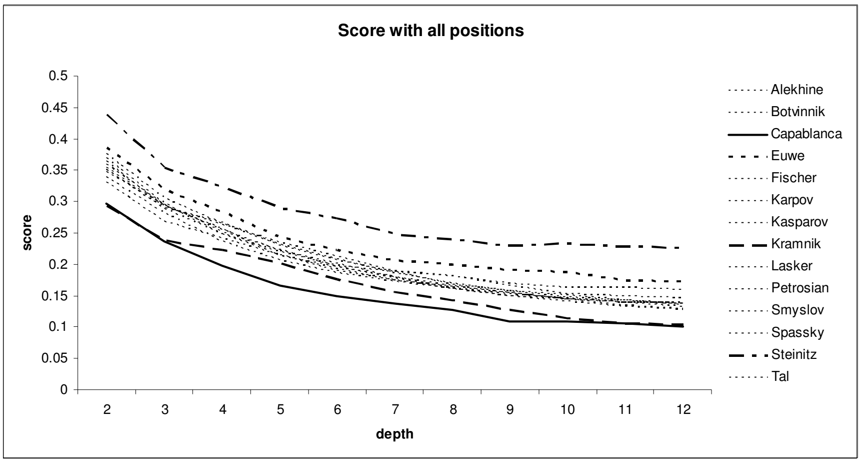


Fig. 3. Average scores (deviations between the evaluations of played moves and best-evaluated moves according to the computer) of each player at different depths of search. The players whose scores clearly deviate from the rest are Capablanca, Kramnik (in positive sense) and Euwe, Steinitz (in negative sense).[4]

The results clearly demonstrate that although the deviations tend to decrease with increasing search depth, the rankings of the players are nevertheless preserved, at least for the players whose scores differ enough from the others (see Fig. 4). It is particularly impressive that even trivial search to depth of two or three ply does rather good job in terms of the ranking of the players.

[4] The lines of the players whose results clearly deviate from the rest are highlighted. The same holds for figures 4, 6, and 8.

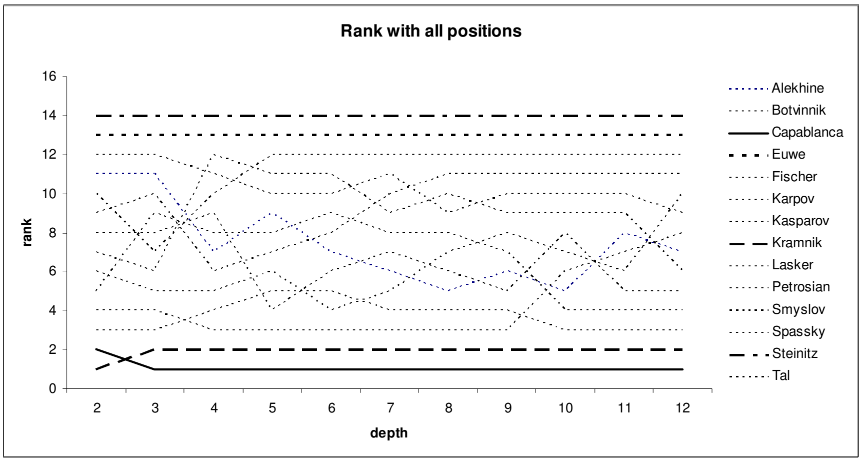


Fig. 4. Ranking of the players at different search depths.

In order to check the reliability of the program as a tool for ranking chess players, it was our goal to determine:

– the stability of the obtained rankings in different subsets of analysed positions,
– the stability of the rankings with increasing search depth.

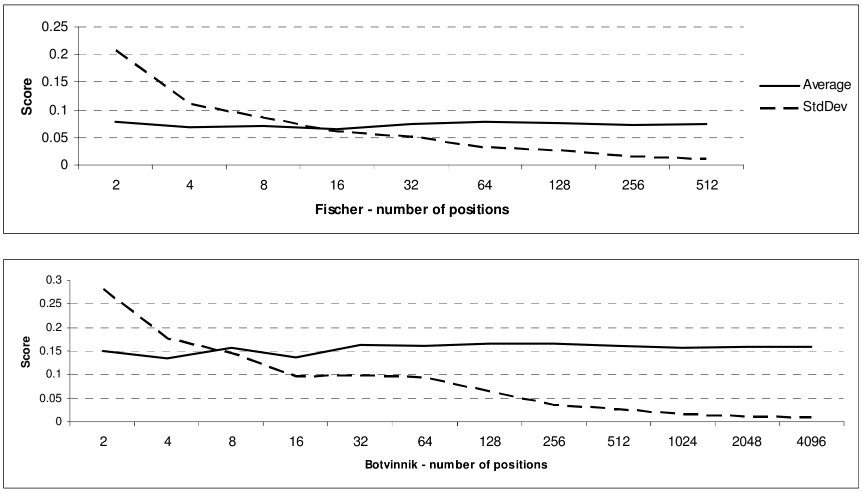


Fig. 5. Average scores of each player were computed for 1000 subsets of different sizes. The graph represents the results for players Fischer and Botvinnik, for subsets consisting of evaluations resulting from search to depth 12.

For each player, 100 subsets from the original dataset were generated by randomly choosing 500 positions (without replacement) from their games. The number of available positions varies for different players, since they were involved in a different number of matches. About 600 positions only were available for Fischer, while both for Botvinnik and Karpov this number is higher than 5100 at each depth. The exact number for each player slightly varies from depth to depth, due to the constraints of the methodology: positions where both the move made and the move suggested by the computer had an evaluation outside the interval [-2, 2] had to be discarded at each depth. Experiments with subsets of different sizes suggest that the size of 500 already seems to be sufficient for reliable results (Fig. 5).

We observed variability of scores and rankings, obtained from each subset, for each player and at each search depth. The results are presented in the next section.

3 Results

The results presented in this section were obtained on 100 subsets of the original dataset, as described in the previous section.

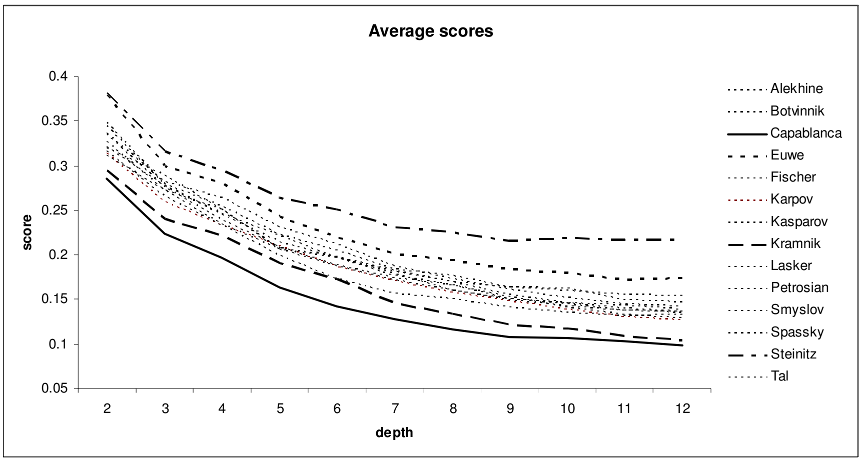


Fig. 6. Average scores of each player at different search depths.

Fig. 6 represents average scores of the players across all the subsets, at each search depth from 2 to 12. The obvious similarity to the graph in Fig. 3 confirms that the results obtained on the whole dataset were not coincidental. This conclusion was confirmed by observing average scores of the players across all depths for each subset separately: Capablanca had the best such score in 96% of all the subsets.

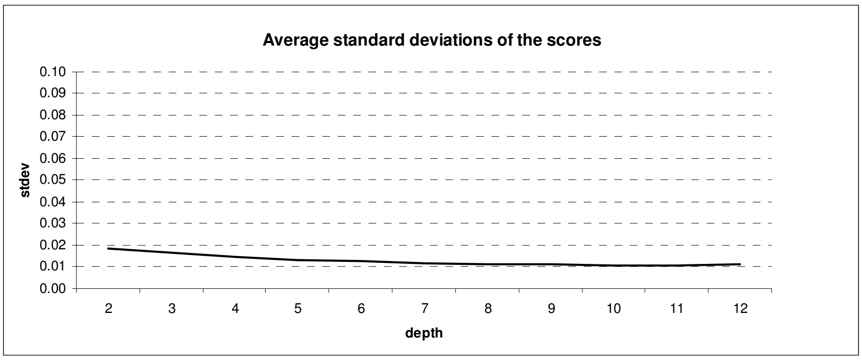


Fig. 7. Average standard deviations of the scores of the players. The scale is adjusted for easier comparison with the graph in Fig. 6.

The average standard deviations of the players' scores show that they are slightly less variable at higher depths. Anyway, they could be considered practically constant at depths higher than 7 (Fig. 7). All the standard deviations are quite low, considering the average difference between players whose score differ significantly.

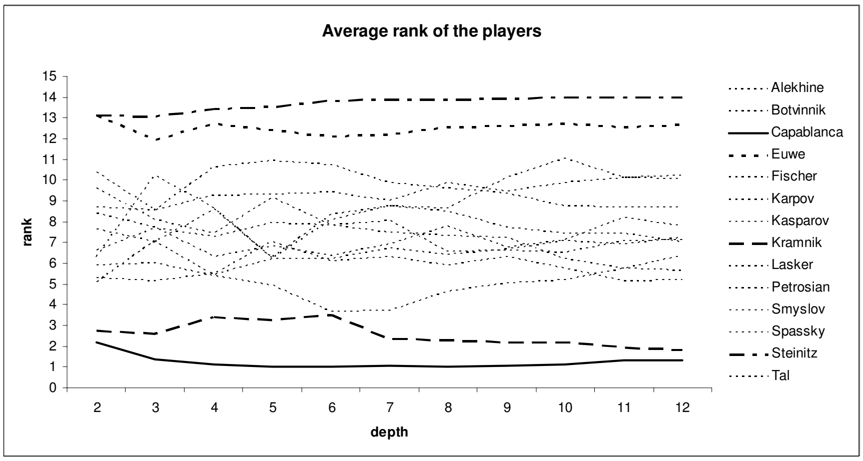


Fig. 8. Average rank of the players.

Fig. 8 (similar to Fig. 4) shows that the rankings preserve for Capablanca, Kramnik, Euwe and Steinitz, whose scores differ significantly from the rest of the players.

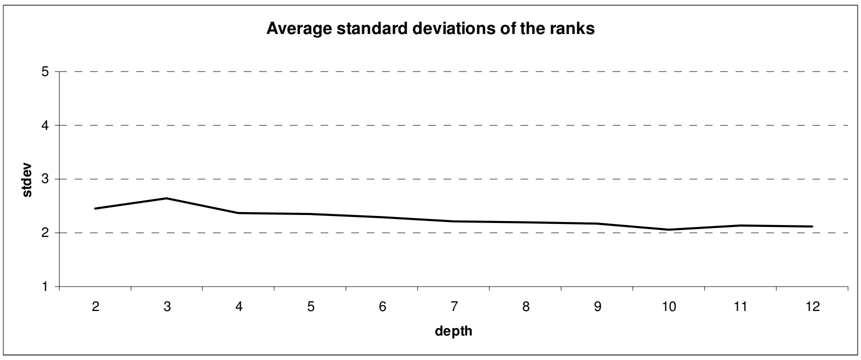


Fig. 9. Average standard deviations of the players' ranks (obtained in 100 subsets).

The average standard deviations of the players' ranks (obtained in 100 subsets) only slightly increase with increasing search depth and are practically equal for most of the depths (Fig. 9).

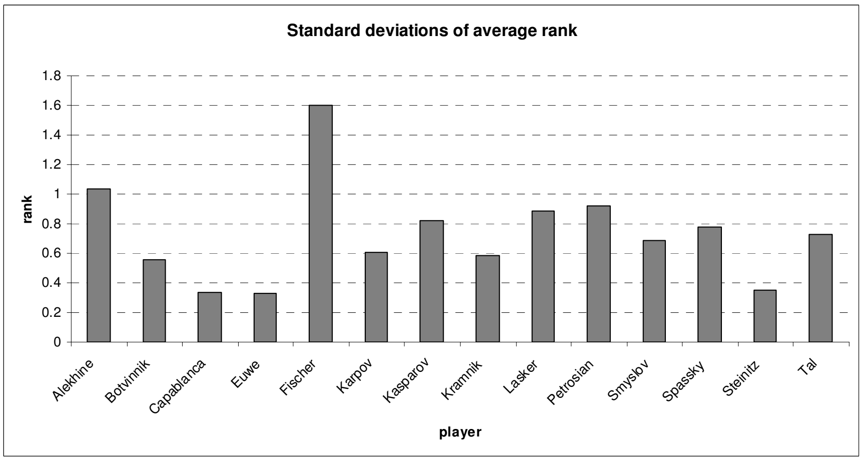


Fig. 10. Standard deviations of the average ranks for each player across all depths.

The graph of standard deviations of the average ranks from different depths for each player separately (Fig. 10) confirms that the rankings of most of the players on average preserve well across different depths of search.

4 A Simple Probabilistic Model of Ranking by Imperfect Referee

Here we present a simple mathematical explanation of why an imperfect evaluator may be quite sufficient to correctly rank the candidates. The following simple model was designed to show the following:

- To obtain a sensible ranking of players, it is not necessary to use a computer that is stronger than the players themselves. There are good chances to obtain a sensible ranking even using a computer that is weaker than the players.
- The (fallible) computer will not exhibit preference for players of similar strength to the computer.

Let there be three players and let us assume that it is agreed what is the best move in every position. Player *A* plays the best move in 90% of positions, player *B* in 80%, and player *C* in 70%. Assume that we do not know these percentages, so we use a computer program to estimate the players' performance. Say the program available for the analysis only plays the best move in 70% of the positions. In addition to the best move in each position, let there be 10 other moves that are inferior to the best move, but the players occasionally make mistakes and play one of these moves instead of the best move. For simplicity we take that each of these moves is equally likely to be chosen by mistake by a player. Therefore player *A*, who plays the best move 90% of the time, will distribute the remaining 10% equally among these 10 moves, giving 1% chance to each of them. Similarly, player *B* will choose any of the inferior moves in 2% of the cases, etc. We also assume that mistakes by all the players, including the computer, are probabilistically independent.

In what situations will the computer, in its imperfect judgement, credit a player for the "best" move? There are two possibilities:

1. The player plays the best move, and the computer also believes that this is the best move.
2. The player makes an inferior move, and the computer also confuses this *same* inferior move for the best.

By simple probabilistic reasoning we can now work out the computer's approximations of the players' performance based on the computer's analysis of a large number of positions. By using (1) we could determine that the computer will report the estimated percentages of correct moves as follows: player *A*: 63.3%, player *B*: 56.6%, and player *C*: 49.9%. These values are quite a bit off the true percentages, but they nevertheless preserve the correct ranking of the players. The example also illustrates that the computer did not particularly favour player *C*, although that player is of similar strength as the computer.

$$P' = P \cdot P_C + (1 - P) \cdot (1 - P_C) / N \qquad \mathbf{(1)}$$

P = probability of the player making the best move
P_C = probability of the computer making the best move
P' = computer's estimate of player's accuracy P

N = number of inferior moves in a position

The simple example above does not exactly correspond to our method which also takes into account the cost of mistakes. But it helps to bring home the point that for sensible analysis we do not necessarily need computers stronger than human players.

5 A More Sophisticated Mathematical Explanation

How come the rankings of the players, as the results demonstrate, preserve rather well, despite the big differences in evaluations across different search depths? In the sequel we attempt to provide an explanation for this phenomenon.

Suppose we have an estimator A that measures the goodness of an individual *M* in a concrete task, by assigning this individual a score (*S*), based on some examples. The estimator assigns different scores to the respective individuals and therefore has a variance associated:

$$Var_M^A = E\left(S_M^A - E\left(S_M^A\right)\right)^2 \qquad \textbf{(2)}$$

The estimator gives an approximation ($S_M{}^A$) of the real score (S_M) of the individual, which results in a bias:

$$Bias_M^A = E\left(S_M^A - S_M\right) \qquad \textbf{(3)}$$

The probability of an error in comparison of two individuals, *M* and *N*, using the estimator *A*, only depends on the bias and the variance. Given two different estimators, *A* and *B*, if their scores are equally biased towards each individual ($Bias_M{}^A$ = $Bias_N{}^A$ and $Bias_M{}^B$ = $Bias_N{}^B$) and variances of the scores of both estimators are equal for each respective individual ($Var_M{}^A$ = $Var_M{}^B$ and $Var_N{}^A$ = $Var_N{}^B$), then both estimators have the same probability of committing an error (Fig. 11).

This phenomenon is commonly known in the machine-learning community and has been frequently used, e.g., in studies of performances of estimators for comparing supervised classification algorithms [1, 5]. In the sequel we analyse what happens in comparisons in the domain of chess when estimators based on CRAFTY at different search depths are used, as has been done in the present paper.

In our study the subscript of $S_M{}^A$ refers to a player and the superscript to a depth of search. The real score S_M could not be determined, but since it is commonly known that in chess the deeper search results in better heuristic evaluations (on average), for each player the average score at depth 12, obtained from all available positions of each respective player, served as the best possible approximation of that score. The biases and the variances of each player were observed at each depth up to 11, once again using the 100 subsets, described in Section 2.

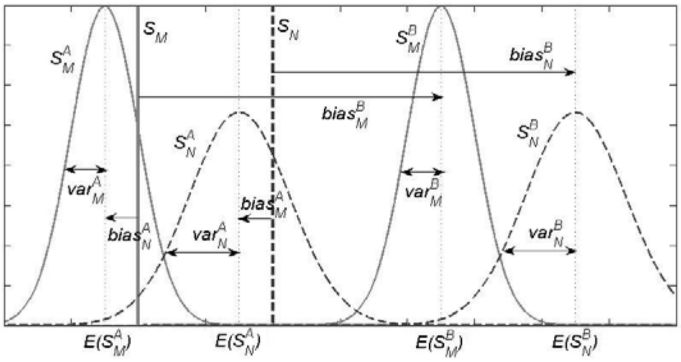


Fig. 11. Although estimators A and B give different approximations of the real scores of individuals M and N (S_M and S_N), and A approximates the real scores more closely, since their scores are equally biased towards each individual ($Bias_M{}^A = Bias_N{}^A$ and $Bias_M{}^B = Bias_N{}^B$) and variances of the scores of both estimators are equal for each respective individual ($Var_M{}^A = Var_M{}^B$ and $Var_N{}^A = Var_N{}^B$), they are both equally suitable for comparison of M and N.

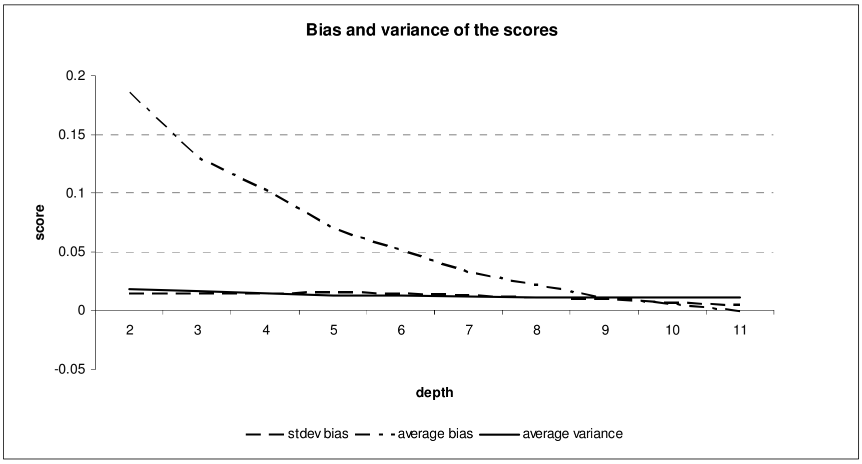


Fig. 12. Average biases, standard deviations of them, and standard deviations of the scores.

The results are presented in Fig. 12. The standard deviation of the bias over all players is very low at each search depth, which suggests that $Bias_M{}^A$ is approximately equal for all the players M. The program did not show any particular bias at any depth towards Capablanca nor towards any other player. Moreover, the standard deviation is practically the same at all levels of search with only a slight tendency to decrease with increasing search depth. Standard deviations of the scores are also very low at all depths, from which we could assume that $Var_M{}^A = Var_M{}^B$ also holds. For better visualisation we only present the mean variance, which as well shows only a slight tendency to decrease with depth. To summarise, taking into account both of these facts, we can conclude that the probability of an error of comparisons performed by CRAFTY at different levels of search is practically the same, and only slightly diminishes with increasing search depth.

6 Conclusion

In this paper we analysed how trustworthy are the rankings of chess champions, produced by computer analysis using the program CRAFTY [4]. In particular, our study was focused around frequently raised reservations expressed in readers' feedback: (1) the chess program used for the analysis was too weak, (2) the depth of the search performed by the program was too shallow, and (3) the number of analysed positions was too low (at least for some players).

The results show that, at least for the two highest ranked and the two lowest ranked players, the rankings are surprisingly stable over a large interval of search depths, and over a large variation of sample positions. It is particularly surprising that even extremely shallow search of just two or three ply enable reasonable rankings. Indirectly, these results also suggest that using other, stronger chess programs would be likely to result in similar rankings of the players.

References

[1] Alpaydin E. Combined 5 x 2 cv F Test for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, Vol.11, No. 8, pp. 1885-1892, 1999.

[2] *Chessbase.com*: Computer analysis of world champions. http://www.chessbase.com/newsdetail.asp?newsid=3465

[3] *Chessbase.com*: Computers choose: who was the strongest player? http://www.chessbase.com/newsdetail.asp?newsid=3455

[4] Guid M. and Bratko I. Computer analysis of world chess champions. *ICGA Journal*, Vol. 29, No. 2, pp. 65-73, 2006.

[5] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (ed. C. S. Mellish), Morgan Kaufmann Publishers, Inc., pp. 1137-1143, 1995.

[6] Steenhuisen J. R. New results in deep-search behaviour. *ICGA Journal*, Vol. 28, No. 4, pp. 203-213, 2005.

[7] *The SSDF Rating List*: http://web.telia.com/~u85924109/ssdf/list.htm

[8] Thompson, K. Computer chess strength. *Advances in Computer Chess 3* (ed. M.R.B. Clarke), Pergamon Press, pp. 55-56, 1982.