HOW TRUSTWORTHY IS CRAFTY'S ANALYSIS OF WORLD CHESS CHAMPIONS?

Matej Guid[1] *Aritz Perez*[2] *Ivan Bratko*[1]

Ljubljana, Slovenia San Sebastian, Spain

ABSTRACT

In 2006, Guid and Bratko carried out a computer analysis of games played by World Chess Champions in an attempt to assess as objective as possible one aspect of the playing strength of chess players of different times. The 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, at least for pairs of the players whose scores differ significantly, it is not very likely that their relative rankings would change if (1) a stronger chess program was used, or (2) if the program would search more deeply, or (3) larger sets of positions were available for the analysis. Experimental results and theoretical explanations are provided to show that, in order to obtain a sensible ranking of the players according to the criterion considered, it is not necessary to use a computer that is stronger than the players themselves.

1. INTRODUCTION

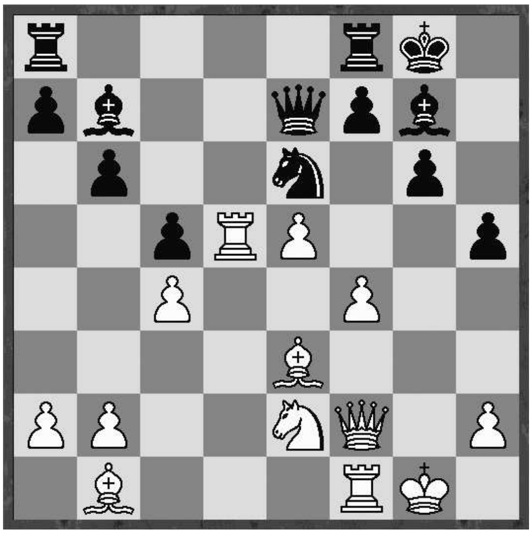
The emergence of high-quality chess programs provided an opportunity of a more objective comparison between chess players of different eras who never had a chance to meet across the board. Recently, Guid and Bratko (2006a) published an extensive computer analysis of World Chess Champions, aiming at such 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 estimate 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. Among several criteria considered, the basic criterion for comparison among players was the average deviation between evaluations of played moves and best-evaluated moves. According to this measure, José Raúl Capablanca, the 3rd World Champion, did best in that analysis. This came as a surprise to many, although Capablanca is widely accepted as an extremely talented and a very accurate player. Of course this is only one of possible measures of performance among many, and this result should be interpreted in the light of Capablanca's playing style that tended towards low complexity positions. Several other criteria were also considered in Guid and Bratko (2006a), e.g., taking into account the playing style of the players and the difficulty of the analysed positions. A method was designed for assessing the complexity of a position. This enabled us to answer questions such as: how would the players under investigation score if they all played in the style of Capablanca or Tal?

Various discussions about the Guid and Bratko (2006a) publication took place at different places, including scientific (Haworth, 2007) as well as popular blogs and forums across the internet.[3] A frequent comment by the readers could be summarised as: "A very interesting study, but it has a flaw in that program CRAFTY, of which the rating is only about 2620, was used to analyse the performance of players stronger than CRAFTY. For this reason the results cannot be useful". Some readers speculated further that the program will give a better ranking to

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

[2] Department of Computer Science and Artificial Intelligence, University of the Basque Country, San Sebastian, Spain.

[3] A slightly different version of Guid and Bratko (2006a) was published by the popular chess website, Chessbase.com (see Guid and Bratko, 2006b). The same website soon published some interesting responses by various readers from all over the world, including some by scientists (Chessbase.com, 2006).



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

Figure 1: Botvinnik-Tal, World Chess Championship match (game 17, position after White's 23^{rd} 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.

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.

1. The program used for analysis was too weak.
2. The depth of the search performed by the program was too shallow.[4]
3. The number of analysed positions was too low (at least for some players).

In this paper, we address these objections in order to assess how reliable CRAFTY (or, by extrapolation, 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 the scores and the rankings of the players are preserved at different depths of search. As Figure 1 illustrates, different search depths may result in large differences in position evaluations. Our results show, possibly surprisingly, that at least for the players whose scores differ sufficiently from the others the ranking remains preserved, even at very shallow search depths.

It is well known for a long time that strength of computer-chess programs increases by search depth. Already in 1982, Ken Thompson compared programs that searched to different search depths. His outcomes showed that searching to only one ply deeper results in a 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 (Steenhuisen, 2005). The preservation of the rankings at different search depths would therefore suggest not only that the same rankings would have been obtained by searching more deeply, but also that using a stronger chess program would probably 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.

We also study in this paper 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 reliable estimates of the players' deviations from the chess program.

To avoid possible misinterpretation of the presented work, it should be noted that this paper is not concerned with the question of how appropriate this particular measure of the playing strength (deviation of player's moves

[4] Search depth in the original study (Guid and Bratko, 2006a) was limited to 12 plies (13 plies in the endgame) plus quiescence search.

from computer-preferred moves) is as a criterion for comparing chess players' ability in general. Therefore any possible interpretations of the results and rankings that appear in this paper should be made carefully keeping this point in mind.

The paper is organised as follows. In Section 2, we describe the methodology used in order to obtain the scores and the rankings of the players. The results of this analysis are presented for each of the players at different search depths. In Section 3, we study experimentally how reliable the results obtained by CRAFTY on available sets of positions are. First, we investigate the question whether the available samples of chess positions were sufficiently large, then we observe the stability of the obtained results repeating the experiments on different subsets of the available positions, and finally we analyse the stability of the obtained scores and rankings across different search depths. Section 4 introduces a simple probabilistic model to show that for a sensible ranking of players, it is not necessary to use a computer that is stronger than the players themselves. Section 5 provides a mathematical explanation of the phenomenon of preservation of rankings regardless of differences in evaluations across different search depths. In Section 6, we discuss some additional aspects of computer analysis of chess players, and summarise the main conclusions of our work.

2. VARIATION WITH SEARCH DEPTH OF DEVIATION BETWEEN PLAYERS AND CRAFTY

In this section we investigate the effects of the search depth on the rankings and the scores of the players, i.e., the average differences between a player's moves and CRAFTY's moves. We used the same methodology as in Guid and Bratko (2006a). Games for the title of World Chess Champion, in which the fourteen classic World Champions contended for or were defending the title, were selected for the analysis. Each position occurring in these games after move 12 was iteratively searched to depths ranging from 2 to 12 ply by the open-source program CRAFTY. The positions before move 12 were excluded to avoid possible effects of chess opening theory. 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 (see Figure 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 (this was done at each depth separately). 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 in this paper 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 study how the rankings of the players fluctuate between different depths of search.

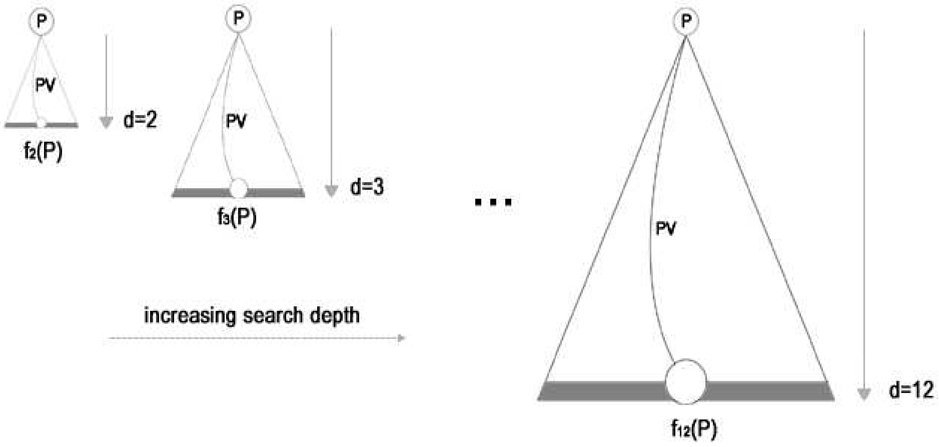


Figure 2: In each position, we performed searches to depths ranging from 2 to 12 plies extended with quiescence search to ensure stable static evaluations. Backed-up evaluations of each of these searches were used for analysis. PV stands for principal variation.

The average differences between the evaluations of the 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. These differences are referred to as players' scores. The score of player *P* at search depth *d* is defined as

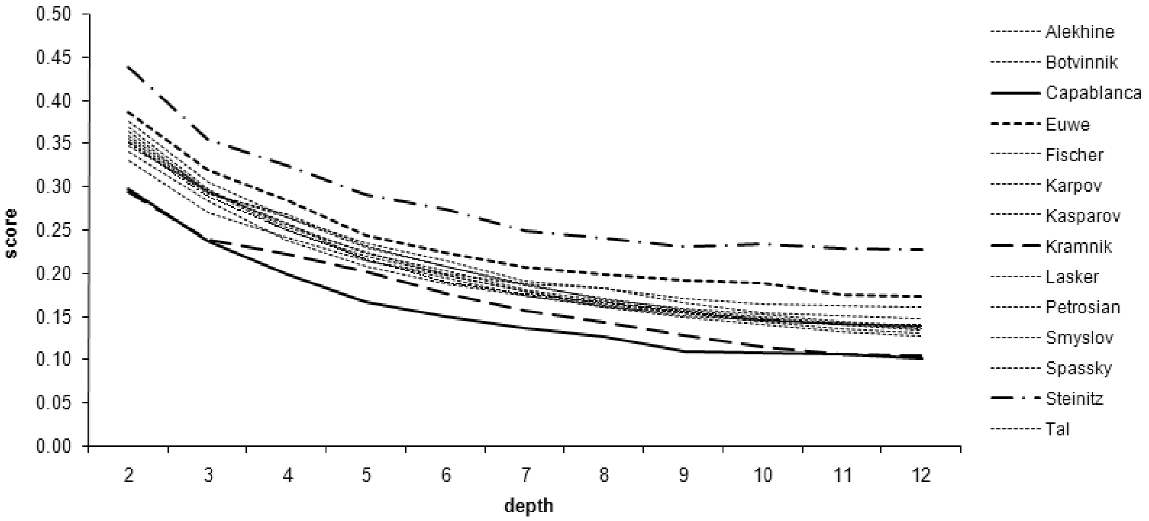


Figure 3: Scores (average deviations between the evaluations of played moves and best-evaluated moves according to CRAFTY) of each player at different depths of search.

$$\frac{\sum |E_{BEST(d)} - E_{PLAYED(d)}|}{N_P(d)} \qquad (1)$$

where $E_{BEST(d)}$ is the evaluation of the move that CRAFTY suggests as best at depth d, $E_{PLAYED(d)}$ is CRAFTY's evaluation of a player's move at depth d, and $N_P(d)$ is the number of moves analysed for player P at particular depth (note that this number varies, since the moves with $E_{BEST(d)}$ and $E_{PLAYED(d)}$ both being outside [-2,2] were discarded). The sum is over all the moves analysed for the player P. Based on players' scores, rankings of the players are obtained in such way that a lower score results in a better ranking.

The players' scores at different search depths are presented in Figure 3, while Figure 4 shows the deviations of the players' scores from the average score of all players obtained at each search depth, and some players whose rankings preserve at most of the depths are highlighted.

The results clearly demonstrate that although the scores of the players tend to decrease with increasing search depth, the rankings of the players are nevertheless preserved at least for the players whose scores differ considerably from the others. It is particularly interesting that even search to depth of just two or three ply (plus quiescence) does a rather good job in terms of the ranking of the players.

3. ROBUSTNESS OF RANKINGS WITH REGARDS TO SAMPLE SIZE

The results presented in the previous section suggest that for some players the obtained rankings are preserved with depth of search. In this section we investigate the question whether the available samples of chess positions were sufficiently large to conclude that the observed differences between pairs of players are statistically significant. We then observe the stability of the obtained results, repeating the experiments on different subsets of the available positions.

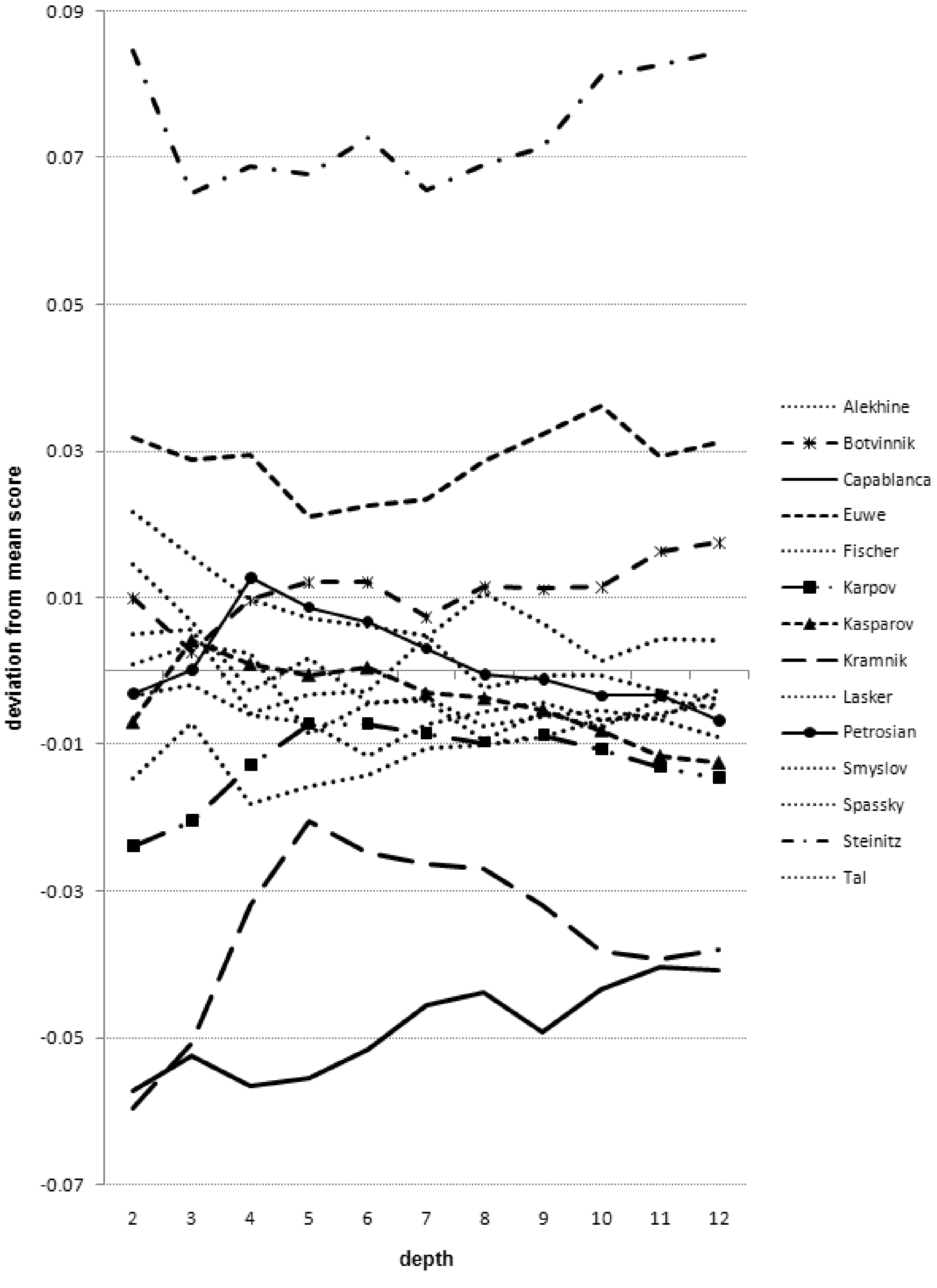


Figure 4: Average deviations of the players' scores from the average score of all players obtained at each depth of search. Based on the players' scores the rankings of the players were obtained. For almost all depths it holds that rank(Capablanca) < rank(Kramnik) < rank (Karpov) < rank(Kasparov) < rank(Petrosian) < rank(Botvinnik) < rank(Euwe) < rank(Steinitz).

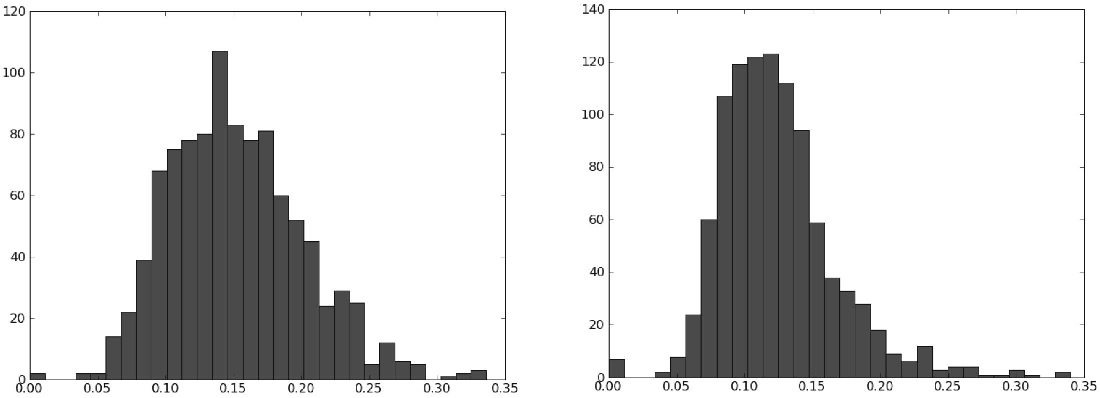


Figure 5: Distributions of scores in 1,000 randomly generated samples consisting of 50 positions are shown for players Fischer (left) and Karpov (right). X axis represents the player's sample score, while Y axis represents the number of samples with the score in a given interval.

3.1 Number of Positions for Analysis

The number of available positions varies for different players. About 600 positions only were available for Fischer,[5] while both for Botvinnik and Karpov this number is higher than 5,000 at each depth. The exact number for each player slightly varies from depth to depth, due to the constraints of the method: positions where both the move made and the move suggested by the computer had an evaluation (based on search to given depth) outside the interval [-2, 2] were discarded at each depth.

To assess whether the set of positions available from World Chess Championship matches were sufficiently large, in order to produce reliable rankings of the players, at least for some pairs of the players, we conducted the following statistical analysis. For each player, n = 30 samples of m = 50 positions were randomly chosen with replacement from the set of all available positions for the player. For each of these positions, we observed the player's deviations from CRAFTY's moves (from now on we will be referring to these deviations as CRAFTY's *differences*) previously computed for a search depth of 12 plies using the method presented in Section 2. For each of the 30 samples, the player's "sample score" was computed as the average of the CRAFTY's differences in the sample.

The sample scores were now used to determine statistical significance of the obtained rankings of the players. Generally speaking, for any two players P_1 and P_2, their mutual rank rank$(P_1) <$ rank(P_2) is determined by the condition score$(P_1) <$ score(P_2). In determining statistical significance, why did we not simply use CRAFTY's differences in the whole data set of individual positions analysed? The reason is that the distributions of the CRAFTY's differences on individual positions are non-symmetrical and very far from normal. Therefore we cannot apply parametric statistical tests on the original data. However, the distributions of the scores (obtained as the average CRAFTY's differences in the sample) in samples consisted of 50 positions are approximately normal (see also the results of an experiment with 1,000 samples of 50 positions given in Fig. 5), so a parametric significance test as the one below is appropriate.

For a pair of players P_1 and P_2, our null hypothesis is that their expected scores are equal. That is, if we had a very large set of positions available for each of them, their observed scores would be indistinguishably close. The alternative hypothesis to the null hypothesis is that the players' expected scores are not equal. Now, given our limited sets of available positions, and the corresponding observed scores and their deviations for the two

[5] In the original study (Guid and Bratko, 2006a) the following candidate matches were included into the analysis in order to compensate for the lack of games of Fischer and Kramnik in their World Chess Championship (WCC) matches: (1) Fischer-Taimanov, Vancouver 1971, (2) Fischer-Larsen, Denver 1971, (3) Fischer-Petrosian, Buenos Aires, 1971, and (4) Kramnik-Shirov, Cazorla 1998. These additional matches were chosen after a careful deliberation: all of them were matches where candidates for the title of World Chess Champion competed under very similar conditions to those in the WCC matches, and all of them took place right before the players contended for the WCC title. In the current study, these matches were not taken into account, and another WCC match was included into analysis: Kramnik-Topalov, Elista 2006 (only the games with slow time control were analysed). This match happened after the results of the original analysis were published. Including this match, Kramnik had more than 1,000 positions available for the analysis.

players, the statistical question is whether the null hypothesis can be rejected at some confidence level, say 95%. If yes, then we may with 95% confidence conclude that the expected scores of the two players are not equal, and therefore the players' performances according to this criterion are not equivalent. We use the following test (for example, see Ross, 2005) to decide this: if the inequality (2) is true then the null hypothesis must be rejected:

$$\sqrt{\frac{s^2_{\bar{X}_1}(m)}{n_1} + \frac{s^2_{\bar{X}_2}(m)}{n_2}} \times z > |M_1 - M_2| \ , \tag{2}$$

where $s^2_{\bar{X}_i}(m)$ is the sample variance of the $n_1 = n_2 = n = 30$ scores of player P_i, $\bar{X}_i$ is a sample score (that is an average CRAFTY's difference in a sample of m ($m = 50$) positions), and M_i is the average of the n sample scores of P_i. The value of z can be obtained from a table of the normal distribution in order to obtain desired confidence levels. Note that a two-tailed test is appropriate for testing our null hypothesis.

We cannot apply this test in our case directly to all pairs of the players, because it is only valid for testing a single hypothesis. Since we have $\binom{14}{2} = 91$ pairs of the players, we need to test 91 hypotheses. Due to the multiple comparisons problem, a more strict test is required. One simple way to strengthen the test is to use the Bonferroni correction where the confidence level is increased by modifying $p = 0.05$ to $p = 0.05/91$, that is dividing p by the total number of tested hypotheses. The Bonferroni correction is however unnecessarily conservative. The False-Discovery-Rate (FDR) method (Benjamini and Hochberg, 1995), a modification of the Bonferroni correction, is a more powerful method which has become popular in many multiple hypothesis testing applications (e.g., Higdon, van Belle, and Kolker, 2008).

Before presenting the results obtained by the FDR method, we look into the question: how large sample sizes m and n can we afford so that the statistical tests are still valid? With increased sample size, more positions are repeated in different samples, so that the samples, which ideally should be independent, may become too similar for reliably estimating the variance. Of course, the smaller the total available set, the more repetitions we have in the samples. And again, the larger the samples are, the more repetition occurs. To check the effect of increased repetitions of positions in the samples of different sizes, we split all Karpov's positions into five subsets of 1,000 positions, and measured the variance on different sized samples drawn from these subsets. We chose Karpov for this experiment because of the large set of positions available from his World Chess Championship games. Finally, we compared the obtained variances with those obtained on the whole set of more than 5,000 positions. Figure 6 shows the results of this experiment. The results indicate that we can afford samples of which the size may be even a rather large proportion of the total set. For example, sample size of 500 out of 1,000 seems perfectly safe. This suggests that our choice of m = 50 was rather conservative.

The results obtained by the statistical test and by using the FDR method for multiple comparisons are shown in Table 1, which shows for which pairs of the players their expected scores differ at the confidence level $>$ 95%. According to the results, for 52.7% of pairs of the players we can with 95% confidence conclude that the expected scores of the two players differ significantly. Note that the sizes of the samples used in our statistical analysis were only m = 50 and n = 30. Therefore these results can be rather conservative.

The results indicate that the sets of available positions were large enough to confirm reliably that for at least one half of the pairs of players their scores differ significantly. So for at least one half of the pairs of the players, their mutual rankings according to chess program CRAFTY would stay the same even if many more positions were available for the analysis.[6] For players whose scores are very similar, however, the positions available from World Chess Championship matches do not produce statistically significant mutual rankings. This indicates that the third reservation posed by some of the readers of the original article (Guid and Bratko, 2006a), speculating that the number of analysed positions was too low (at least for some players), should be taken seriously, at least when looking for a firm statistical guarantee regarding the relative rankings of some pairs of the players.

3.2 Stability of the Rankings with Search Depth

The results presented in the sequel were obtained on 100 subsets of the original data sets, generated by randomly choosing 500 positions (with replacement) from the available position samples of each player. The aim here is to study the stability of rankings across the search depths.

[6] Of course, this assumes that positions selected for computer analysis appropriately represent the strength of a particular player. In Guid and Bratko (2006a) it is explained why the games from WCC matches were selected as representative for each particular player.

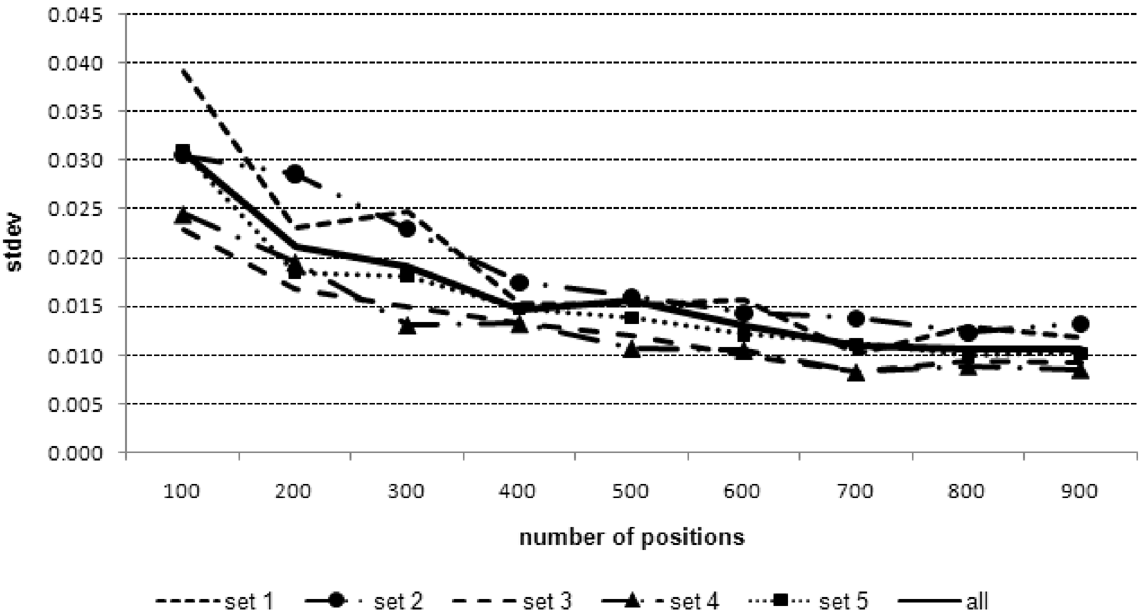


Figure 6: Standard deviations of the scores in 100 subsets of positions (n = 100) of different sizes m that were obtained on 5 data sets that each consisted of 1,000 positions from Karpov's games. Different positions were included in each dataset. The results that were obtained from all available positions from Karpov's games are included as well.

In order to study the stability of the scores at different samples, the standard deviation of the scores at different search depths were obtained for each of the players. The results are summarised in Figure 7, which shows the averages of the obtained standard deviations. The average standard deviations of the players' scores show that they are less variable at higher depths. Anyway, they could be considered practically constant at depths higher than 7. We also observed that Capablanca had the best score in 95% of all the subset-depth combinations.

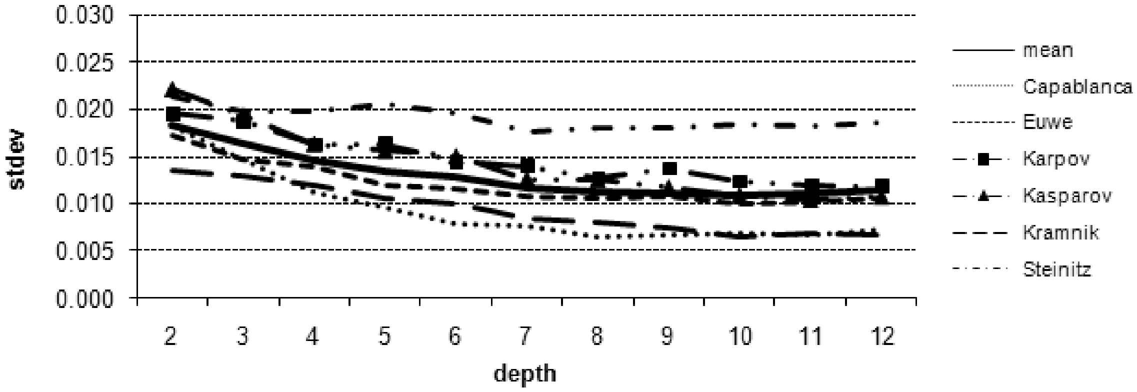


Figure 7: Average standard deviations of the scores of the players over 100 random subsets of 500 positions, and standard deviations of the scores of some of the players (for clarity, only a few players are included).

In order to determine the stability of the rankings (obtained in 100 subsets) across different search depths, standard deviations of the ranks of individual players at each search depth were computed. The results are summarised in Figure 8, which shows the average of the standard deviations of all the players. They only slightly decrease with increasing search depth. Yet, we may state that they are practically equal for most of the depths (see Figure 8).

	Ste	Euw	Bot	Tal	Las	Fis	Smy	Ale	Pet	Spa	Kas	Kar	Kra	Cap
Capablanca	X	X	X	X	X	X	X	X	X	X	X	X	X	
Kramnik	X	X	X	X	X	X			X	X	X	X		
Karpov	X	X												
Kasparov	X	X												
Spassky	X	X												
Petrosian	X	X				X								
Alekhine	X	X	X			X								
Smyslov	X	X	X											
Fischer	X	X												
Lasker	X	X												
Tal	X	X	X											
Botvinnik	X													
Euwe	X													
Steinitz														

Table 1: The names of the players are ordered according to their scores at search depth 12 that were obtained on the whole set of positions available from their World Chess Championship matches. Pairs of the players whose expected scores differ at the confidence level > 95% are marked with 'X'. The results were obtained by the false discovery rate (FDR) procedure for multiple comparisons.

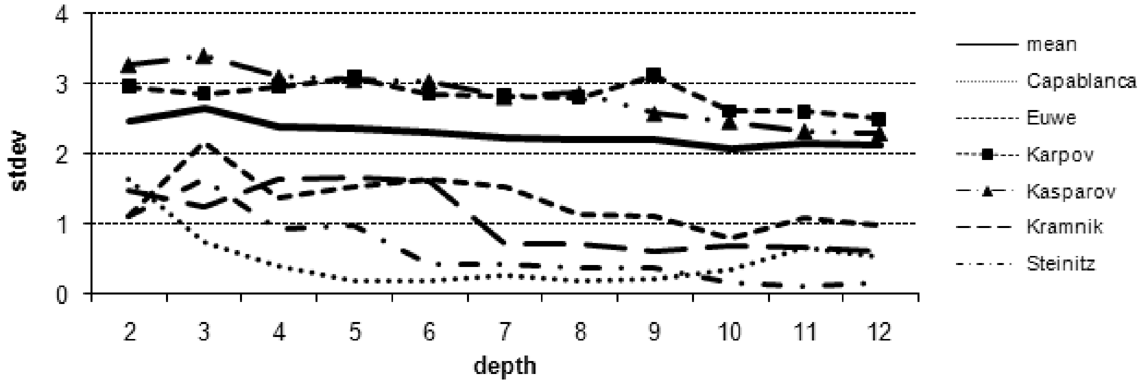


Figure 8: Average standard deviations of the players' ranks (obtained in 100 subsets), and standard deviations of the ranks of some of the players (for clarity, only a few players are included).

Finally, we observed the stability of the obtained ranks for each player across different search depths, i.e., how much do the players' ranks tend to change at different search depths. The results of this study are summarised in Figure 9, which shows standard deviations of the average ranks for each player across all the search depths. The low standard deviation values for most of the players (lower than 1) confirm 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 AN IMPERFECT REFEREE

Here, we present a simple mathematical explanation of why an imperfect referee (also called evaluator) may be sufficient to rank the candidates correctly. The following simple model was designed to show two points.

1. 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 when using a computer that is weaker than the players.
2. The (fallible) computer will not exhibit preference for players of similar strength to the computer.

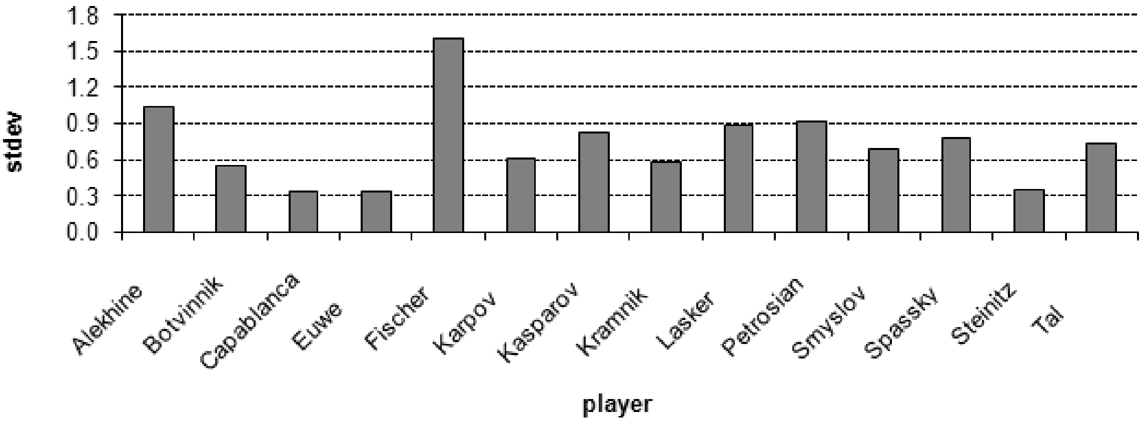


Figure 9: Standard deviations of the average ranks for each player across all depths.

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 the inferior 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.

In general in this model, the computer's estimate of a player's accuracy can be calculated as follows. Let

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

Then:

$$P' = P \times P_C + \frac{(1-P) \times (1-P_C)}{N} \tag{3}$$

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 equation (3) we can 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 (i.e., 90%, 80%, and 70% for players A, B, and C respectively), 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.

The straightforward 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.

P' is monotonically increasing with *P* as long as $P_C > 1$ / (*N*+1). Note that P_C corresponds to random referee in the case when P_C = 1 / (*N*+1). So according to this model, the referee only has to be better than random to obtain the ranking right, given (1) sufficiently large samples of positions, and (2) the independence assumption being true. That is, the computer's choice of wrong moves is independent of the player's wrong moves. All this is not to say that a perfect referee and a referee just better than random are equally useful in determining rankings. In a realistic setting, where position sets are limited, an inferior referee is more likely to obtain the ranking wrong because of larger statistical fluctuations in smaller samples.

5. VARIANCE OF PLAYERS' SCORES AND RANKINGS WITH SEARCH DEPTH

Assume we have an estimator *A* that measures the performance of an individual *M* at a concrete task, by assigning this individual a score *S*, based on some examples of *M* performing the task. The estimator assigns different score values to the individual at different examples, and the associated variance and bias are:

$$Var^A_M = E[(S^A_M - E(S^A_M))^2] \tag{4}$$

$$Bias^A_M = E(S^A_M - S^A_M) \tag{5}$$

Moreover, assuming a normal distribution of score values, the probability of an error in the relative rankings 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 ($Bias^A_M$ = $Bias^A_N$ and $Bias^B_M$ = $Bias^B_N$) and variances of the scores of both estimators are equal for each respective individual (Var^A_M = Var^B_M and Var^A_N = Var^N_B), then both estimators have the same probability of committing an error (see Figure 10). 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 (for example, see Kohavi (1995)).

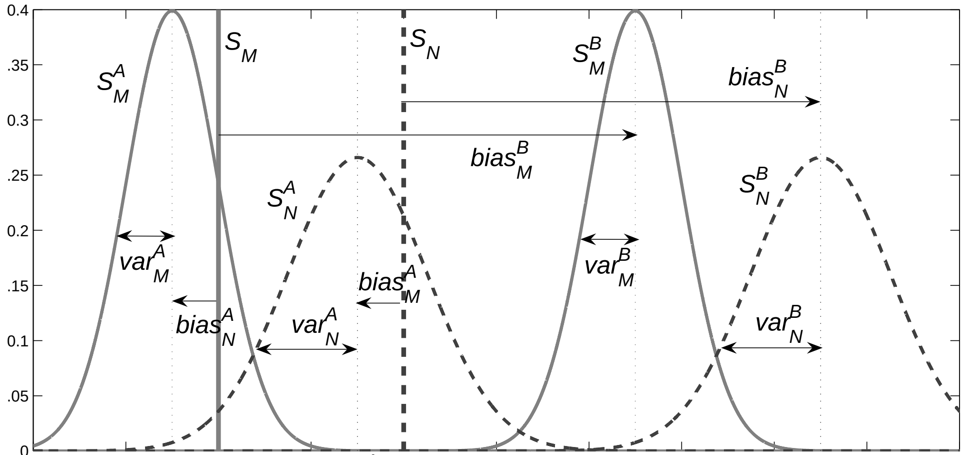


Figure 10: Although estimators *A* and *B* give different approximations of the true performances of individuals *M* and *N* (S_M and S_N), and *A* approximates the real scores more closely, since their scores are equally biased towards each individual ($Bias^A_M$ = $Bias^A_N$ and $Bias^B_M$ = $Bias^B_N$) and variances of the scores of both estimators are equal for each respective individual (Var^A_M = Var^B_M and Var^A_N = Var^B_N), they are both equally suitable for mutual ranking of *M* and *N*.

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 M in S_M^A refers to a player and the superscript A to a depth of search. The true performance 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 score at depth 12, obtained from all available positions of each respective player, served as the best possible approximation of the true performance. The biases and the variances for each player were observed at each depth up to 11, once again using the 100 subsets of 500 positions, described in Section 3.2.

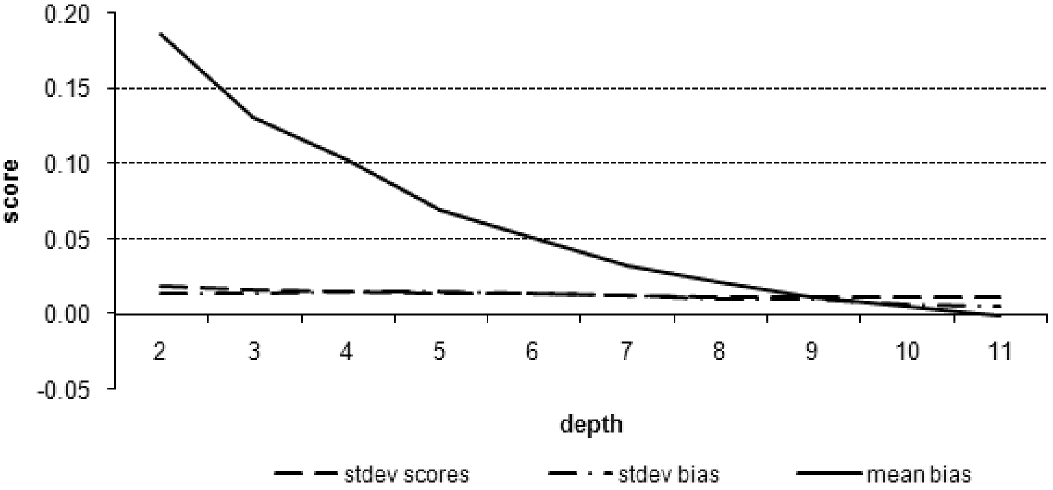


Figure 11: Average biases, standard deviations of them, and standard deviations of the scores with 100 subsets.

The results are presented in Figure 11. 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 neither show any particular bias at any depth towards Capablanca nor towards any other player, if we assume that CRAFTY at search depth 12 is not biased. From Figure 11 we make two observations. (1) The standard deviation is practically the same at all levels of search with only a slight tendency to decrease with increasing search depth. (2) Standard deviations of the scores are very low, too, at all depths, from which we may infer that Var_M^A = Var_M^B also holds. For a 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 observations, we may 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. DISCUSSION AND CONCLUSIONS

In this paper, we analysed how trustworthy the scores and rankings of chess champions are, when produced by computer analysis using the program CRAFTY (see Guid and Bratko, 2006a). In particular, our study was focussed around three possible problems with this analysis: (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). A brief summary of the conclusions regarding these three possible problems is: (1) the chess program used is unlikely to be too weak, (2) the depth of search is unlikely to be too low, and (3) the number of analysed positions was sufficient to demonstrate statistical significance of the differences in the scores between more than one half of the pairs of 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 position sample. Even extremely shallow search of just two or three ply enable reasonable mutual rankings for some pairs of the players. Indirectly, the results of this work also suggest that using other, stronger chess programs would be likely to result in similar rankings of the players whose scores differ by more than average margin.

Statistical analysis of the results shows that for at least one half of the pairs of the players the differences in their scores are statistically significant at 95% confidence or higher. This result was obtained with strict test that takes into account the number of tested statistical hypotheses.

Last but not least, our experimental findings strongly suggest 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.

One frequent question by the readers was associated with the meaning of the players' scores obtained by the program. A typical misinterpretation of their meaning went like this: "For every 8 moves on average, CRAFTY expects to gain an advantage of one extra pawn over Kasparov" (Chessbase.com, 2006). We would like to emphasize here that the scores obtained by the program only measure the average differences between the players' choices of move and the computer's choice. The experimental results presented in this paper demonstrate that the scores are associated with the strength of the program and are not invariable for the same program at different depths of search. However, as the analysis shows these scores that are relative to the computer used, have good chances to produce sensible rankings of the players. The decreasing tendency of the scores with increasing search depth also suggests that the scores obtained by a program stronger than CRAFTY would be lower than the scores obtained by CRAFTY.

For appropriate interpretation of the obtained scores and rankings of the players, it should be emphasized again that this is only one possible criterion for the comparison of the players among many sensible criteria of very different kinds. This paper is concerned only with the credibility of the obtained results (Guid and Bratko, 2006a) in the estimates according to this particular criterion, by studying the three critical questions mentioned above. From the chess player's point of view, this criterion is particularly crude in that it does not take into account the differences in the average difficulty of the positions played by different players. Nor does this score-based criterion take into account another important aspect, that is the differences between the playing styles of different players.

There are many other questions of interest that were not addressed in this paper, including: (1) How to take into account the differences between players in the average difficulty of the positions encountered in their games; (2) Does CRAFTY's style of play exhibit preference for the styles of any particular players? These two questions have been studied in Guid and Bratko (2006a). However, we recommend further work on these questions.

7. ACKNOWLEDGEMENT

This work was partly supported by the Slovenian research agency ARRS. We would like to thank the anonymous referees for many useful comments. Janez Demšar, Martin Možina, Aleksander Sadikov, Lan Umek, Gorazd Lampič, and Matej Marinč helped with advice and discussion of various aspects of this paper. Finally, we also thank the many readers of the original paper (Guid and Bratko, 2006a) for the large number of interesting points among their ample feedback.

8. REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1, pp. 289–300.

Chessbase.com (2006). Computers choose: who was the strongest player? http://www.chessbase.com/newsdetail.asp?newsid=3465.

Guid, M. and Bratko, I. (2006a). Computer Analysis of World Chess Champions. *ICGA Journal*, Vol. 29, No. 2, pp. 65–73.

Guid, M. and Bratko, I. (2006b). Computer Analysis of World Chess Champions. http://www.chessbase.com/newsdetail.asp?newsid=3455. Chessbase.com.

Haworth, G. (2007). Gentlemen, Stop Your Engines! *ICGA Journal*, Vol. 30, No. 3, pp. 150–156.

Higdon, R., Belle, G. van, and Kolker, E. (2008). A note on the false discovery rate and inconsistent comparisons between experiments. *Bioinformatics*, Vol. 24, No. 10, pp. 1225–1228.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *14th International Joint Conferences on Artificial Intelligence (IJCAI 1995), Proceedings*, pp. 1137–1143, Morgan Kaufmann, Los Altos, CA.

Ross, S. M. (2005). *Introductory Statistics.* Elsevier Academic Press.

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

Thompson, K. (1982). Computer Chess Strength. *Advances in Computer Chess 3* (ed. M. R. B. Clarke), pp. 55–56, Pergamon Press, Oxford, UK.