

Computer analysis of World Chess Champions

Matej Guid and Ivan Bratko

University of Ljubljana,
Faculty of Computer and Information Science,
Artificial Intelligence Laboratory,
Tržaška 25, 1000 Ljubljana, Slovenia

Abstract. Who is the best chess player of all time? Chess players are often interested in this question that has never been answered authoritatively, because it requires comparison between chess players of different eras who never met across the board. In this paper, we attempt such a comparison based on the evaluation with a chess playing program of games played by the world chess champions in their championship matches. We slightly adapted the program Crafty for this purpose. Our analysis also takes into account the differences in players' styles to account for the fact that calm positional players have in their typical games less chance to commit gross tactical errors than aggressive tactical players. To this end, we designed a method to assess the difficulty of positions. Some of the results of this computer analysis might appear quite surprising. Overall, the results can be nicely interpreted by a chess expert.

1 Introduction

Who was the best chess player of all time? This is a frequent and interesting question, to which there is no well founded, objective answer, because it requires comparison between chess players of different eras who never met across the board. With the emergence of high quality chess programs a possibility of such an objective comparison arises. Despite this fact, computers were so far mostly used as a tool for statistical analysis of players' results. However, such statistical analyses often do not reflect true strengths of the players, nor do they reflect their quality of play. It is common that chess players play against opponents of different strengths and also that the quality of play changes in time. Furthermore, in chess a single bad move can decisively influence the final outcome of a game, even if all the rest of the moves are excellent. Therefore, the same result can be achieved through play of completely different quality.

The most complete and resounding attempt made to determine the best chess player in history has recently been put forward by Jeff Sonas, who has become a leading authority in the field of statistical analysis in chess during past years. Sonas devised a specialized rating scheme, based on tournament results from 1840 to the present [1]. The rating is calculated for each month separately, with player's activity taken into account. A player's rating, therefore, starts declining when he is no longer active, which differs from the classic FIDE rating.

Having a unified system of calculating ratings represents an interesting solution to determining a "common denominator" for all chess players. However, it does not take into account that quality of play has risen dramatically in the recent decades. The first official world champion, Steinitz, achieved his best Sonas rating, which is on par with ratings of recent champions, in April 1876. His rating is determined from his success in tournaments in time when the general play quality was well below that of today. The ratings in general reflect the players' success in competition, but not directly their quality of play.

Other estimates about who was the strongest chess player of all times, are based primarily on the analyses of their games done by chess grandmasters and these are often subjective. Thirteenth World Chess Champion, Gary Kasparov, in his unfinished set of books *My Great predecessors* [2], analyses in detail numerous games of best chess players in history and will most probably express his opinion regarding who was the best chess player ever. But it will be merely *an opinion*, although very appreciated in the chess world.

Our approach was different: we were interested in the chess players' quality of play regardless of the game score, which we evaluated with the help of computer analyses of individual *moves* made by each player.

2 Method

We evaluated fourteen classic version world champions, since the first World Chess Championship in 1886 to the present. Matches for the title of "World Chess Champion", where they contended for or were defending the title, were selected for analysis.

Roughly, the basis for evaluation of a human's play was the difference between position values resulting from the moves played by the human and the moves chosen as best by the chess program. This approach can be criticized on the basis that sometimes there are alternative, equally strong moves, and the choice between them is the matter of playing style and not merely chess strength. We will return to this issue later and provide a refinement and a justification for this approach.

Evaluation of each game started on the 12th move, without the use of an openings library, of course. This decision was based on the following careful deliberation. Not only today's chess programs poorly evaluate positions in the first phase of a game, but also analysing games from the start would most likely favour more recent champions, due to vast progress made in the theory of chess openings. Starting the analyses on a later move would, on the other hand, discard too much information. The chess program Crafty [3], which we slightly modified for the purpose of analyses, was used. Instead of time limit, constant *fixed search depth* was applied on every move. With such an approach we achieved the following:

- Complex positions, which require processing larger search trees to achieve a more accurate evaluation, automatically get more computation time.

- The program could be run on different computers without fear of not getting the same evaluations for a given set of positions on each of them.

The latter enabled us to greatly speed up the calculation process by distributing the computation among a network of machines, and as a consequence, a greater search depth was possible. We chose to limit search depth to 12 plies plus quiescence search. There were some speculations that a program searching 12 plies would be able to achieve a rating that is greater than that of the World Champion [4], arguably a long time ago. However, the search depth mentioned was chosen as the best alternative, since deeper search would mean a vast amount of additional computation time (more than ten full days of computation time on 36 machines with an average speed of 2.5 GHz were required to perform analyses of all games). The limit of search was increased to 13 plies in the endgame. Crafty’s definition of endgame was used – it starts when the total combined numerical value of both white and black pieces on board (without pawns) is less than 15. We also changed the ‘king’s safety asymmetry’ parameter thus achieving a shift from Crafty’s usual defensive stance to a more neutral – where it was neither defensive nor offensive. Quiescence search remained turned on to prevent horizon effects.

With each evaluated move, data was collected for different search depths (which ranged from 2 to 12), comprising of best evaluated move and the evaluation itself, second best evaluated move and its evaluation, the move made by the human and its evaluation. We also collected data on material state of both players from the first move on.

2.1 Average difference between moves made and best evaluated moves

The basic criterion was the average difference between numerical evaluations of moves that were played by the players and numerical evaluations of moves that were suggested by computer analysis as the best possible moves.

$$\text{MeanLoss} = \frac{\sum |\text{best move evaluation} - \text{move played evaluation}|}{\text{number of moves}} \quad (1)$$

Additional limitations were imposed upon this criterion. Moves, where both the move made and the move suggested had an evaluation outside the interval $[-2, 2]$, were discarded and not taken into account in the calculations. The reason for this is the fact that a player with a decisive advantage often chooses not to play the best move, but rather plays a move which is still ‘good enough’ to lead to victory and is less risky. Similar situation arises when a player considers his position to be lost – a deliberate objectively worse move may be made in such a case to give the player a higher practical chance to save the game against a fallible opponent. Such moves are, from a practical viewpoint, justified. Taking them into account would wrongly penalise players that used this legitimate approach trying (and sometimes succeeding) to get to a desired result. All positions with evaluations outside the interval specified were declared lost or won.

2.2 Blunders

Big mistakes or blunders can be quite reliably detected with a computer, to a high percentage of accuracy. Individual evaluations could be inaccurate, but such inaccuracies rarely prevent the machine from distinguishing blunders made in play from reasonable moves.

Detection of errors was similar to the aforementioned criterion – we used a measure of difference between evaluations of moves played and evaluations of moves suggested by the machine as the best ones. We label a move as a blunder when the numerical error exceeds 1.00, which is equivalent to losing a pawn without compensation. Like before we discarded moves where both evaluations of the move made by a player and the move suggested by the machine lie outside $[-2, 2]$ interval, due to reasons already mentioned.

2.3 Complexity of a position

Deficiency of the two criteria, detailed in the previous sections, is in the fact that there are several types of players with specific properties, which the criteria do not account for. It is reasonable to expect that *positional players* in average commit less errors due to somewhat less complex positions in which they find themselves as a result of their style of play, than *tactical players*. The latter on average deal with more complex positions, but are also better at handling them and use this advantage to achieve excellent results in competition.

We wanted to determine how players would perform when faced with equally complex positions. In order to determine this, a comparison metric for position complexity was required.

Although there are enormous differences in the amount of search, nevertheless there are similarities regarding the way chess programs and human players conduct a search for the best possible move in a given position. They both deal with a giant search tree, with current position as the root node of the tree, positions that follow with all possible moves as children of the root node, and so on recursively for every node. They both search for best continuations and doing so they both try to discard moves that are of no importance for evaluation of the current position. They only differ in the way they discard them. A computer is running algorithms for efficient subtree pruning whereas a human is depending mainly on his knowledge and experience. Since they are both limited in time, they cannot search to an arbitrary depth, so they eventually have to evaluate a position at one point. They both utilize partial evaluations at given depths of search. While computer uses evaluations in a numerical form, human player usually has in mind descriptive evaluations, such as “small advantage”, “decisive advantage”, “unclear position”, etc. Since they may have a great impact on the evaluation, they both check all forced variations (the computer uses *quiescence search* for that purpose) before giving an assessment to the root position. One can therefore draw many parallels between machine and human best move search procedures, which served as a basis for assessing the complexity of positions.

The basic idea is as follows: a given position is difficult with respect to the task of accurate evaluation and finding the best move, when different “best move”, which considerably alter the evaluation of the root position, are discovered at different search depths. In such a situation, a player has to analyse more continuations and search to a greater depth from the initial position to find moves that may greatly influence the assessment of the initial position and then eventually choose the best continuation.

As complexity metric for an individual move, we chose the sum of absolute differences between the evaluation of the best and the second best move, invoked every time a change in evaluation occurs when search depth is increased. A corresponding algorithm for calculating the complexity of positions is:

```
complexity := 0

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

The difference between the evaluations of the best and the second best move represents the significance of change in the best move when search depth is increased. It is reasonable to assume that a position is of higher complexity, that is more difficult to make a decision on a move, when larger changes regarding the best move are detected when increasing search depth. Merely counting the number of changes of the best move at different search depths would give an inadequate metric, because making a good decision should not be difficult in positions where several equally good choices arise.

We used the described metric of position complexity to determine the distribution of moves played across different intervals of complexity, based on positions that players found themselves in. This in turn largely defines their *style of play*. For example, Capablanca who is regarded as a calm positional player, had much less dealing with complex situations compared to Tal, who is regarded as a tactical player. For each player that was taken into consideration, the distribution over complexity was determined and average error for each complexity interval was calculated (numerical scale of complexity was divided into intervals in steps of 0.1). We also calculated an average distribution of complexity of moves made for the described intervals for all players combined.

The described approach enabled us to calculate an expected average error of world champions in a hypothetical case where they would all play equally complex positions. We calculated the errors for two cases. Firstly, for a game of average complexity, averaged among games played by all players and secondly for

a game of average complexity, averaged among games played by a single player. The latter represents an attempt to determine how well the players would play, should they all play in the style of Capablanca, Tal, etc.

2.4 Percentage of best moves played and the difference in best move evaluations

The percentage of best moves played alone does not actually describe the quality of a player as much as one might expect. In certain types of position it is much easier to find a good move than in others. Experiments showed that the percentage of best moves played is highly correlated to the difference in evaluations of the best and second best move in a given position. The greater the difference, the better was the percentage of player's success in making the best move (Fig. 1).

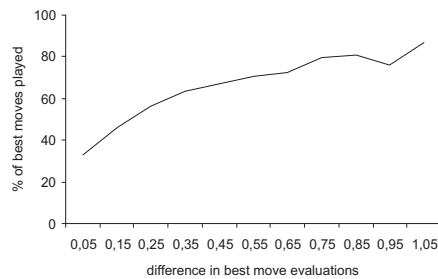


Fig. 1. Proportion of best moves played in dependence of difference in best move evaluations

Such a correlation makes sense, because the bigger the difference between the best two moves, the greater the error made when selecting the wrong move. The height of the curve is amplified by the fact that we are dealing with world champions, experts at the game of chess. Analysis of weaker players would give a curve of lesser height.

By analysing the correlation between the percentage of best moves played and the difference in best two move's evaluations, we derive information about the quality of each individual player. It turned out that curves for individual players differ significantly. This behaviour served as basis for creating a criterion, used to infer information on quality of individual players.

For each player we calculated the distribution of moves across separate intervals of the difference in evaluations of two best moves (where the step was 0.1). We also calculated an average distribution for all players combined. Given this average distribution, we then determined the expected percentage of best moves played for each individual player. Due to reasons already mentioned, we did not count clearly lost or won positions in this statistics.

2.5 Material

The purpose of calculating average material quantity, that is the sum of numerically expressed values of all pieces on board, was not to determine the quality of play, but to collect additional information on player's style of play. Mainly we tried to observe player's inclination to simplify positions.

2.6 Credibility of Crafty as an analysis tool

It was important to determine whether Crafty represents a valid analysis tool for evaluating world champions. Chess programs of the present time are mostly being regarded as weaker than the best human chess players. It is very likely that Crafty is weaker than at least some of the world champions that were taken into consideration.

There are many arguments in favour of computer programs being an appropriate tool for evaluating chess players: they use numerical values as evaluations, they adhere to the same rules all the time and are therefore more consistent than human observers. In particular they are very good at evaluating tactical positions, where a lot of computation is required.

Modified Crafty that was used in our work, has a great advantage when compared to standard chess programs. By limiting and fixing search depth we achieved automatic adaptation of time used to the complexity of a given position. Another important fact is that we were able to analyse a relatively large sample of 1397 games, containing over 37.000 positions – therefore occasional errors made in the evaluating of positions have little affect on the final, averaged results.

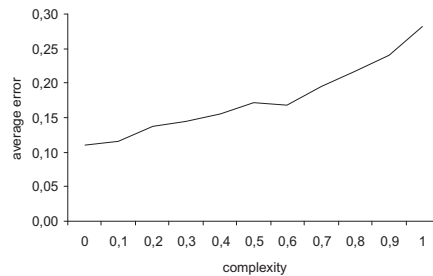


Fig. 2. Graph of errors made by players at different levels of complexity clearly indicates the validity of the chosen measure of complexity of positions; the players made little errors in simple positions, and the error rate increased with increasing complexity

To assess how trustworthy Crafty is as our assumed gold standard, we checked the correlation between our calculated error rates made in the games and the actual outcomes of these games. As we stated, in our opinion game results do not always reflect the actual quality of play and mere statistical analysis of game

outcomes is not enough to compare world champions. Because of this, we did not expect absolute correlation, but for Crafty’s credibility a significant level of correlation should be detected nonetheless. We determined the correlation between the difference in measured average errors made by opposing players in a given game and the outcome of that game. Calculated Spearman correlation was found to be $\rho = 0.89$ (with significance level $p < 0.0001$).

3 Results

The basic criterion for evaluating world champions was the average difference between moves played and best evaluated moves by computer analysis.

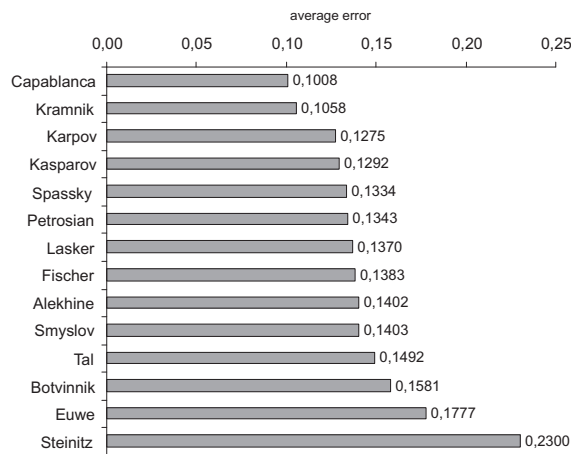


Fig. 3. Average difference between moves played and best evaluated moves (loss per move)

According to this analysis (Fig. 3), the winner was the third world champion, Jose Raul Capablanca. We expected positional players to perform better by this criterion than tactical players. Capablanca is widely renowned to be a pure positional player. On the other hand Steinitz, who lived in an era of tactical “Romantic chess”, took clearly last place.

The results of blunder rate measurement are similar (Fig. 4). Notice the excellent result of Petrosian, who is widely renowned as a player who almost never blundered. Gary Kasparov describes Capablanca with words “He contrived to win the most important tournaments and matches, going undefeated for years (of all the champions he lost the fewest games).” and “his style, one of the purest, most crystal-clear in the entire history of chess, astonishes one with his logic.” [2]

Capablanca is renowned for playing a “simple” chess and avoiding complications, while it is common that Steinitz and Tal faced many “wild” positions

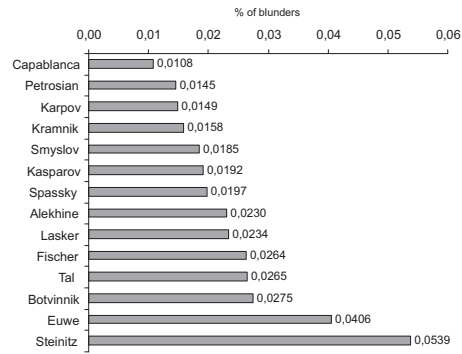


Fig. 4. Blunder rate

in their games. The results of complexity measurement (Fig. 5) clearly coincide with this common opinion.

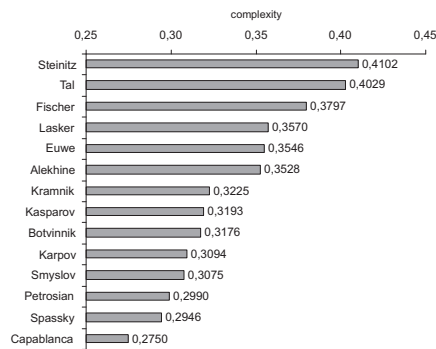


Fig. 5. Average position complexity

Fig. 6 demonstrates that Capablanca indeed had much less dealings with complex positions compared to Tal. Distribution of moves in different intervals regarding complexity has a lot to do with player's *style*. Calculated players' expected errors with various such distributions was another criterion. The winner was the fourteenth world champion Vladimir Kramnik. Kramnik also had the best performance of all the matches – his average error in his match against Kasparov (London, 2000) was only 0.0903. It is interesting to notice that Kasparov would outperform Karpov, providing they both played in Tal's style.

Another criterion was expected number of best move played providing that all players dealt with positions with equal difference between the best two moves, as was described in the previous section. It represents another attempt to bring

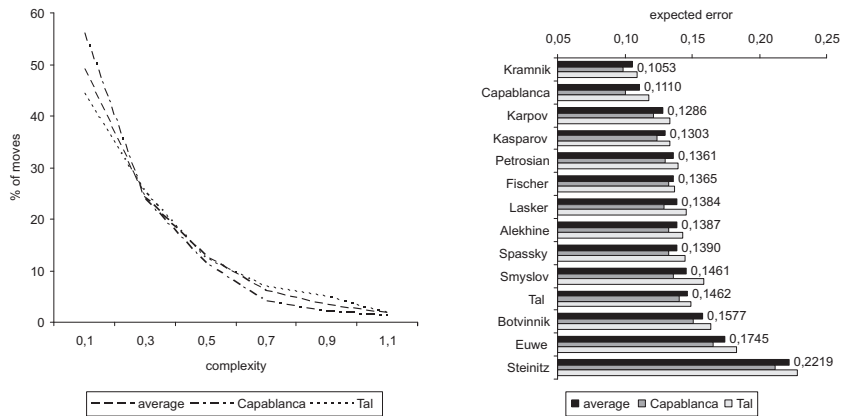


Fig. 6. Expected errors when playing in different styles

the champions to a common denominator. Kramnik, Fischer and Alekhine had the highest percentage of best moves played, but also the above mentioned difference was high. On the contrary, Capablanca, who was right next regarding the percentage of best move played, on average dealt with the smallest difference between the best two moves. The winner by this criterion was once again Capablanca. He and Kramnik again clearly outperformed the others.

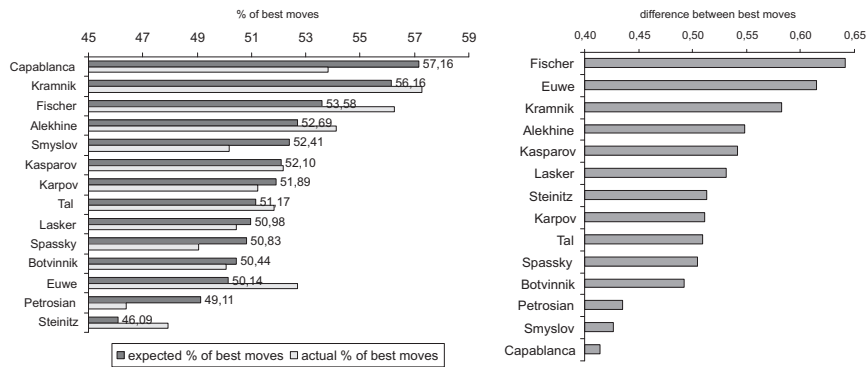


Fig. 7. Percentage of best move played and the difference between best two moves

The graphs in Fig. 8 show players' tendencies to exchange pieces. Among the players who stand out from the others, Kramnik obviously dealt with less material on board. The opposite could be said for Steinitz, Spassky and Petrosian.

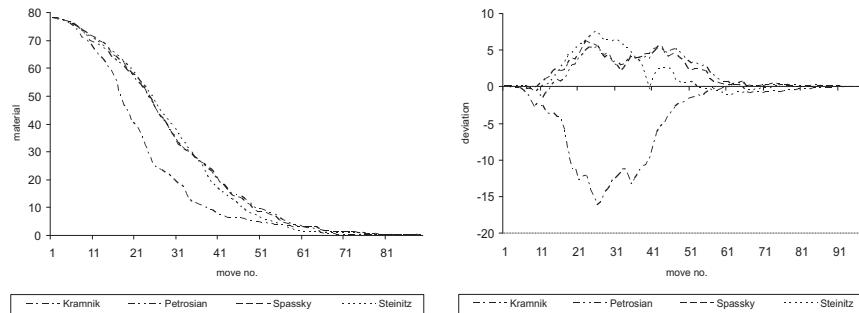


Fig. 8. Material during the game and players' deviations regarding it

4 Conclusion and future work

We applied the slightly modified chess program Crafty as tool for computer analysis of games played by world chess champions aiming at an objective comparison of chess players of different eras. Generally, the results of our computer analysis can be nicely interpreted by a chess expert. Some of the results might appear quite surprising and may thus be considered also as an interesting contribution to the field of chess. Capabalanca's outstanding score in terms of mean value loss will probably appear to many as such an interesting finding, although it probably should not come as a complete surprise. As we did in this study, this result should be interpreted in the light of the comparatively low complexity of positions in Capabalanca's games which is quite in line with the known assessments in the chess literature of his style. For example, Kasparov [2] when commenting Capabalanca's games speculates that Capabalanca occasionally did not even bother to calculate deep tactical variations. The Cuban simply preferred to play moves that were clear and positionally so strongly justified that calculation of variations was simply not necessary.

Our approach assumes that Crafty's evaluation, based on search limited to 12 ply plus quiescence, is accurate enough to be used as the golden standard. It seems indeed that this worked fine in our analysis. Even if Crafty's evaluations are not always perfect, for our analysis they just need to be sufficiently accurate on average since small occasional errors cancel out through statistical averaging. Still, as one idea for future work, it would be nice to obtain some more firm, quantitatively supported evidence about evaluation error of Crafty with respect to some sort of ideal evaluation.

A related question is whether using more recent chess programs that in tournaments perform better than Crafty would make a significant difference if applied instead of Crafty. This question is difficult to answer directly by simply plugging another program into the analysis system instead of Crafty, because these other programs would have to be modified for the analysis similarly as Crafty. That would require source code of these programs that was not available. An

indirect way of tentatively answering this question is however possible by evaluating these strong chess programs by our method using Crafty. High scores of these programs evaluated by Crafty would indicate that Crafty competently appreciates the strength of these programs, and that thus using these programs to evaluate human players instead of Crafty would be likely to produce similar results. To retain the style of human play, we chose to use for this experiment games played between these top programs against top human players. Evaluation by Crafty of strong programs gave the following results: Deep Blue's mean loss per move was 0.0757 (in 6 games match with Kasparov, New York 1997), Deep Fritz mean loss 0.0617 (8 games with Kramnik, Bahrain 2002), Deep Junior mean loss 0.0865 (6 games with Kasparov, New York 2003), Fritz X3D mean loss 0.0904 (4 games with Kasparov, New York 2003), Hydra 0.0743 (6 games with Adams, London 2005). These results give some indication that using other strongest chess programs instead of Crafty would probably not affect the results significantly.

As mean evaluation loss per move is obviously not sufficient to assess a player's strength, we also took into account the average difficulty of positions encountered in the player's games. This made it possible to compare players of different playing styles. Our measure of position complexity seems to have produced sensible results. These results are qualitatively much in line to how an expert chess commentator would describe the players in this study in terms of their playing style. As another line of future work, it would be interesting to explore by means of a psychological study, how well our complexity measure reflects the true cognitive difficulty of a chess position.

References

1. Sonas, J.: Chessmetrics. <http://www.chessmetrics.com> (2005)
2. Kasparov, G.: My Great Predecessors, Parts 1-5. Everyman Chess (2003-2006)
3. Hyatt, R.: The crafty ftp site. <ftp://ftp.cis.uab.edu/pub/hyatt/> (2006)
4. Hsu, F., Anantharaman, T., Campbell, M., Nowatzyk, A.: A grandmaster chess machine. *Scientific American* **263**(4) (1990) 44-50