THE BRATKO-KOPEC EXPERIMENT: A COMPARISON OF HUMAN AND COMPUTER PERFORMANCE IN CHESS

D. Kopec and I. Bratko*

Machine Intelligence Research Unit, University of Edinburgh, UK

ABSTRACT

Recently the best computer programs have demonstrated the ability to hold their own against grandmasters in blitz play and in tournament play have been able to obtain ratings just below the master level. The foundation of their success is the ability to exhaustively search 6 to 7 or more ply which makes them superior in tactical positions to humans of the same rating but not necessarily in positional play.

We have designed the experiment in order to obtain some quantitative support for this proposition. Our positions have been chosen with the view that a certain type of positional move (called a lever) can play an important role in the strong player's ability to find the best move in a position. Our hypothesis is that strong computer programs will score better than humans of the same rating on tactical problems but rather more poorly on the selected positional problems, unless the best positional move also leads to gain within their search limits.

INTRODUCTION

Computer and Human Chess

It has been our long held view that in artificial intelligence work, particularly with regard to computer chess, more attention should be paid to the way humans do things before attempting to implement the computational process involved. De Groot's (1965) work with chess masters established the fact that they build small lookahead trees, generally storing about 30 positions in their lookahead memory, with an upper bound of the order of 100 positions. This leads to the conclusion that a chess master's unique talent

*Present address: Faculty of Electrical Engineering and E. Kardelj University, Joseph Stefan Institute, Ljubljana, Yugoslavia

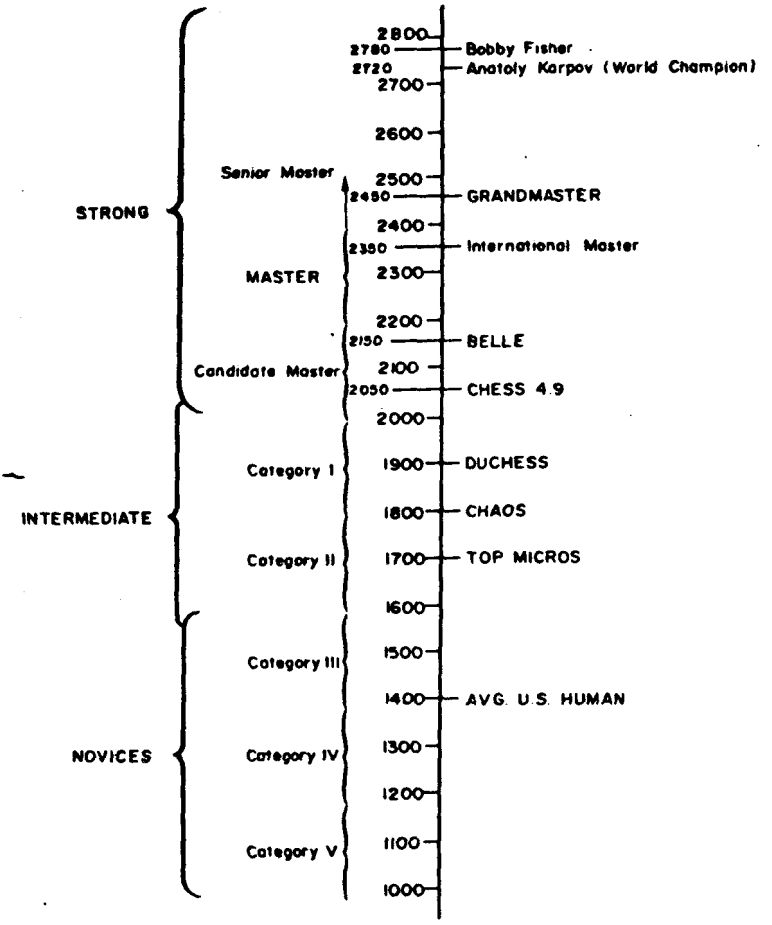

Fig. 1 The Elo Rating System with the three major categories used for purposes of this experiment on the left, U S categories in the middle, and some key points where humans and certain computer chess programs fall on it.

categories: tactical or positional. Tactical moves are those which involve the interaction and possible capture of White and Black forces and include

(1) checkmate or gain of material
and/or (2) a distinct improvement in terms of positional ends (e.g. mobility)
and/or (3) the defence to some immediate threats.

Positional moves are those which do *not* involve interaction of the opposing White and Black forces but result in improvements in such tangible notions as mobility, centralisation, acquisition of new terrain (space or squares), regroupment of forces, etc.

Related Work

Four further works provide the spirit and background of our present research. E.T.O. Slater (1950) recorded the differences in mobility between winners and losers of 78 arbitrarily selected master games which ended in a decisive result on or before the 40th move; see Michie (1980). This helped to establish the importance of *mobility* which is still employed as a significant factor in the evaluation function of most modern computer chess programs. Tan's (1977) work pointed towards the complexities of pawn endings and attempted to develop a logical framework which might uncover their secrets. The vocabulary for Tan's work is that defined in *Pawn Power in Chess* (Kmoch, 1959). One term in particular provides the motivation behind our present experiment, *levers*. Kmoch's simple overall definition is as follows (p. 16): "The situation in which two opposing pawns can capture each other constitutes an element of pawn play which we shall call the *lever*...". Our definition includes a few additions though the overall concept is unaltered: a pawn move which

(1) offers to trade itself,
(2) leads to an ultimate improvement in the pawn structure of the side playing it
and/or (3) damages the opponent's pawn structure.

This is founded on the notion that any pawn structure can be reliably defined and measured in terms of positive and negative points. An example of a lever which results in the improvement of the pawn structure of the side playing it is given in Figure 2.1, while a lever of the type which damages the opponent's pawn structure is given in Figure 2.2.

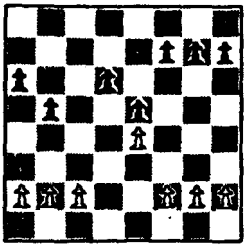

Fig. 2.1 A classic Sicilian Defence pawn structure whereby if Black can safely play the lever ...d5 he gets rid of a weakness and improves his pawn structure.

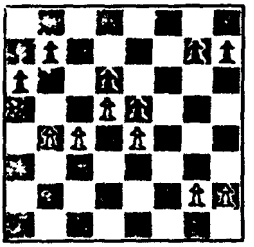

Fig. 2.2 A position where a lever of the second type, 1 c5, damages Black's d6-e5 mini-chain and forms a duo.

Levers may be considered as a subclass of positional moves, though they may also be considered to border on the realm of tactical chess in the sense that they involve the interaction of opposing forces and may employ a temporary or long-term pawn sacrifice.

More than two centuries ago Philidor (1749) said: "Les pions sont l'ame du jeu" – pawns are the soul of chess. They are even more: they provide the "skeleton", the overall concept or outline of a position. An effort to establish their role in the strong player's ability to recall a position was earlier work by Bratko, Tancig and Tancig (1976), where very strong players were tested on their ability to recall a set of stimulus positions from short term memory. They found that the pawns were recalled much more consistently than other pieces, particularly when organized into some well-known patterns. They also found that the ability to recall the positions of pieces was directly related to how well they fit into these pawn configurations or patterns (see also Chase and Simon, 1973). This is the foundation of our decision to use lever moves for the choice of those experimental positions in which the correct move is a positional one.

THE BRATKO-KOPEC EXPERIMENT

Pilot Experiment

Our original experiment in 1977 consisted of 25 stimulus positions, 20 of which were 'lever' positions from *Pawn Power in Chess* with five tactical positions included as controls. The positions were stored in a data file on the DEC10 at Edinburgh's Regional Computing Centre and then flashed for 1½ minutes each on a hard-wired chess TV display unit. Subjects were then allowed 30 seconds to write down their choice of 'best move(s)' and 'candidate move(s)' for each position. Only human chessplayer subjects in Edinburgh were tested. Our general finding was that scores correlated closely with ratings and that with some experience we were soon able to predict subjects' scores a priori, based on their ratings. Where scores were higher than would be expected from subjects' ratings, we have found that their subsequent substantial improvement in rating had been effectively foreshadowed. To draw attention to 'bias' we asked each subject to note after the experiment whether he had read *Pawn Power in Chess*. However, the number (5) of control tactical positions proved insufficient to draw any conclusions on the relative roles played by tactics and levers according to a player's rating. It was also difficult to standardize the scoring of candidates moves by our experimental design. These findings enabled us to conclude that in further experimentation it would be necessary to

(1) increase the number of tactical positions;
(2) substitute the notion of 'candidate moves' with '2nd choice', '3rd choice', '4th choice';
(3) make the experiment more portable.

The Experimental Design

Of the 20 original lever positions from *Pawn Power in Chess*, ten were retained with two additional ones selected from *The Best Move* (Hort and Jansa, 1980), and nine additional tactical positions were chosen from *Informator 18* (Matanovic, 1975) and *Modern Chess Tactics* (Pachman, 1973), with three of the original five being retained. Thus 24 positions (12 tactical, denoted by T, and 12 levers denoted by L) are presented on the

separate pages of a booklet with the side to move indicated in brackets after the identifying number of each diagrammed position as well as on a standardized answer sheet. Subjects were given a total of two minutes for each position to select their preferred move(s) and to write down up to four choices in order of preference on the answer sheet provided. Thus the experiment is portable and can be administered, e.g. by mail, to any chess-player, human or machine, in the world.

Results

Human Subjects

Thus far we have tested 35 human chess-player subjects and 12 computer chess-playing programs. Scoring 1/N where N goes from one to four as the choice-preference of the correct move(s). So that if the "preferred move" selected by a subject for a given test position is the correct move, one full point credit is given; if the subject's second choice is the correct move, then 1/2 point credit is given; third choice correct gives 1/3 point credit, and fourth choice gives 1/4. The distribution of subjects' scores within six rating zones according to their dependence on T and L is given in Table 1, and a graph of score against rating in Figure 3.

TABLE 1

Rating Range	Mean T	Mean L	Mean 12(T-L)/S	Number of subjects
1000–1599	1.13	1.05	0.43	8
1600–1799	3.33	4.13	-1.32	4
1800–1999	5.43	4.62	0.97	7
2000–2199	6.67	4.84	1.73	6
2200–2399	7.81	8.07	0.20	8
2400 +	10.50	9.95	0.32	2
Overall mean	5.23	5.09	0.39	Total 35

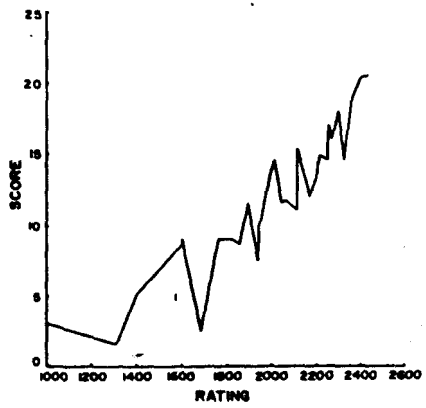

Fig. 3 Graph of rating vs score (35 humans)

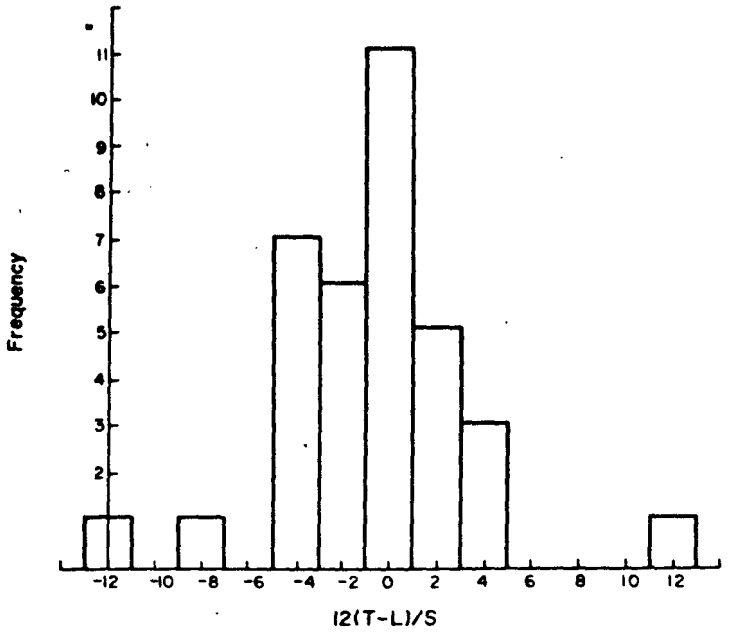

Fig. 4.1 Histogram of 12(T-L)/S for 35 human subjects (T = tactics, L = lever, S = total score)

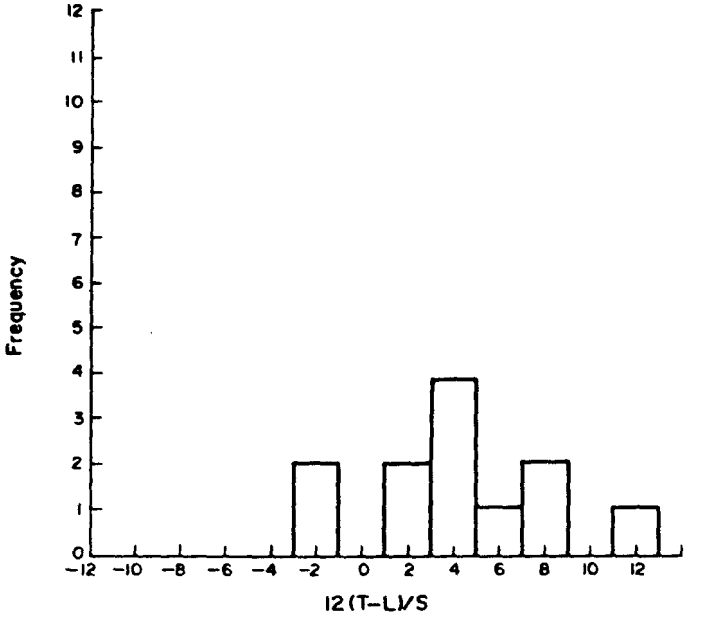

Fig. 4.2 Histogram of 12(T-L)/S for 12 computer chess programs

A further refinement of the differences between these T and L scores and their relative effect on total score (S) is given by the histograms in Figures 4.1 and 4.2. Here the ratio (T-L)/S, the proportional deviation, has been calculated and then multiplied by 12 for scaling purposes. It is worth noting that in the case of very high scoring subjects the (T-L)/S ratio has its bounds. For example, if a subject scores 18, the maximum for T > L is T = 12, L = 6, so that 12(T-L)/S gives 4. This means that simply due to the experimental design in that the maximum of both T and L is 12, high scores (i.e. > 18) are comprised of general success with regard to both T and L. Therefore a greater difference between T and L in scores near 12 is most significant, as is the case with the group rated 2000-2199. However, a more appropriate measure would be to use the ratio (T-L)/S if S is less than or equal to 12 and to use the ratio (T-L)/(24-S) if S is greater than 12. A further 15 sample subjects have not been included here due to various unreliablity factors such as age, rating, etc.

Computer Subjects

Though our data for computer programs as compared to humans is somewhat limited, Table 2 indicates that for scores over 5 there is a strong trend for T > L. The two cases where L > T (3 to 2) can be attributed to the fact that the scores are low and possibly to the nature of at least one of the positions involved. This will be discussed in the next section.

TABLE 2 Computer Subjects

	Program	Rating	Score	T	L	12(T-L)/S
1.	Chess Challenger '10'	Unr	1	1	0	+12.0
2.	Chess Challenger '7'	Unr	5	2	3	- 2.4
3.	Sensory Chess Challenger	Unr	5	3	2	+ 2.4
4.	Sargon 2.5	1720~	5	2	3	- 2.4
5.	AWIT	1400	5	4	1	+ 7.2
6.	OSTRICH81	1450~	6	4	2	+ 4.00
7.	CHAOS	1820	6	5	1	+ 8.0
8.	Chess Champion MK V (E)	1885~	6.83	5	1.83	+ 5.56
9.	Morphy Encore	1800~	9.33	6	3.33	+ 3.43
10.	BCP	1685~	13	10	3	+ 6.46
11.	DUCHESS	1850	16.50	10.5	6	+ 4.38
12.	BELLE	2150	18.25	11	7.25	+ 2.46

Key: (E) Experimental version; ~ Rating is an estimate.
Note: Programs running off mainframe computers have names entirely in upper case letters. Others are stand-alone microcomputer programs.

Discussion of Positions

The set of 24 positions is given as an appendix, as is the "Master Sheet" which gives the correct move(s) in each position and the sources. We refer to positions by their number followed by the side to move in brackets. [At this point the reader might like to go to the appendix and try the test for herself before reading further. *Ed.*]

14(W) is straightfoward tactics; 1 Qd2 or 1 Qe1 wins heavy material.

15(W) is from a Fischer game which many subjects, particularly younger ones, recognised. After 1 Qxg7+ Qxg7 2 Rxf6 Qxg3 3 hxg3 later followed by g4-g5-g6, Fischer managed to trade off his extra doubled P to remain a P up.

16(W) is an example of a tactical position whereby after 1 Ne4! White is guaranteed at least positional gains with 2 Nd6+ to follow; i.e. if 1 ...dxe4 2 Bxf7+ Kxf7 3 Qxd8 hxg5, though Black obtains three pieces for his Q, his exposed K, P-deficit, and lack of piece co-ordination mean that he does not have sufficient compensation. However, after 1 ...Be6 (as suggested in BCP's search) 2 Nd6+ etc., White only obtains a big positional plus.

17(B) calls for 1 ...h5 with the idea of ...hg, Nh7 and Ng5 to follow. If 2 g5 Nh7 3 h4 f6!. Alekhine played 1 ...Ne8 (a move suggested by many subjects) in this position and did not obtain good play.

18(B) is from a Fischer game which exemplifies the fact that the achievement of the two bishops versus bishop and knight in a semi-open position is at the highest level tantamount to material gain. Very few humans found 1 ...Nb3, most stronger ones suggesting 1 ...Qb6 or 1 ...Be6. After 1 ...Nb3 2 Bxb3 Qb6+ White relinquishes the two bishop advantage to Black and is left weakened on the light squares. The programs BELLE and DUCHESS found 1 ...Nb3.

19(B) is the "fork trick" in action. After 1 ...Rxe4 2 Rxe4 d5 3 Qxa6 dxe4 4 Be3 Qg4! Keres managed to transfer his central advantage to a winning K-side attack.

20(W) suggests the straightforward lever 1 g4 with the intention to follow 1 ...fg with 2 Qxg4 and f5, striking at the base of Black's chain and exposing his disorganised position.

In 21(W) 1 Nh6 wins the exchange in all variations.

22(B) is the hardest position of the entire set, at least for humans. Perhaps the fact that only one human subject, International Master Craig Pritchett, found the best move as did the programs BCP, DUCHESS and BELLE is most significant to the experiment. Humans suggest reasonable and/or interesting moves such as Rfd8, Nc5, d5!?, Ne5!? and Nh5, which often come into consideration in similar positions, but the most unusual combination starting with 1 ...Bxe4 followed by 2 ...Qxc4 is the key. It should be noted that depth of search is not the problem for humans in finding this combination; but rather more likely is its individuality and the fact that many good moves seem in the offing.

23(B) is also a hard position in the sense that the "normal" move 1 ...Bf5 is confronted with the very interesting 2 g4!? which most people (and machines) fail to consider adequately. 1 ...f6 is an indisputable, solid lever which meets the threat 2 f5.

Finally, in 24(W) 1 f4 is the indicated lever since White's superior pieces make it easier for him to maintain the tension in the centre.

A number of human subjects made interesting comments and criticisms after participating in the experiment. Some suggested that they would have fared much better had they been given an initial few "training" positions to get some idea of what was being asked for in the experiment. However, this would give us no fair method of comparing human results with computer results. Others stated that in a number of positions they could guess the "characteristic" move we were after though in two minutes or under tournament time constraints (rather over 2 minutes per move) they could not calculate its

consequences and would most probably not play the indicated move. Quite a few subjects, particularly the younger ones, recognized the Fischer position, 15(W), where he played Qxg7+ against Mecking in the 1970 Palma de Mallorca Interzonal. Nevertheless we do not feel that this or indeed any other position that may have been recognized invalidates their inclusion in the experiment. A chess player's experience or education can be used as a measure of his ability. We accept that a few positions in the experiment are not ideal, and that a few are even controversial as to what the best move is, but this will not significantly invalidate a human or machine subject's overall score when compared with the standard deviation.

CONCLUSIONS

The design of the experiment facilitates the quantitative study of differences in the ways that human players and most tournament programs play chess. The results confirmed those differences, which were suspected prior to the experiment. This confirmation is particularly obvious in the T/L diagram of Figure 5. It is hoped that the experiment can become a standard test for the characteristics of chess programs by enabling the establishment of their "tactic vs lever" profile.

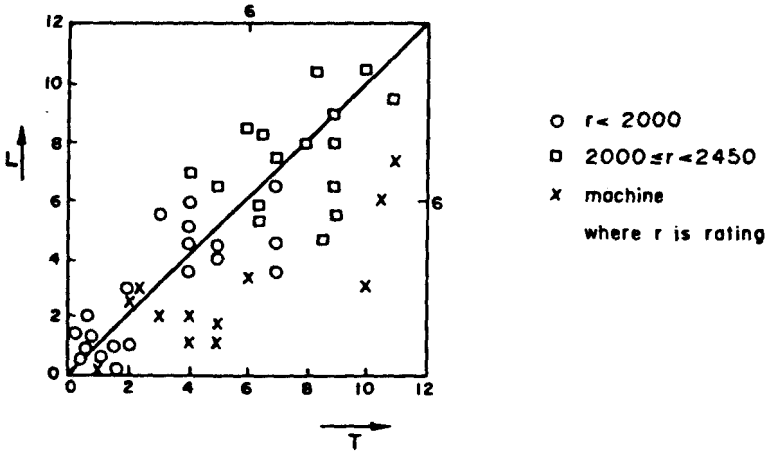

Fig. 5 T/L profile for humans and machines. Note that machines largely predominate in the T zone below the diagonal.

A few comments should be made about the disadvantages of the experiment. The first is based on the observation that some of the programs scored surprisingly well, outscoring strong human players who in our judgement would be able to beat these programs under standard tournament conditions. One explanation for this is that the test conditions were more favourable to machines than humans. Human players tend during actual games to non-uniformly allocate their total time to individual moves. Thus a chess master would typically spend 10 or 20 minutes or more in a critical position for finding a key move or a correct plan, and then play the next few moves almost

instantly. On the contrary, most programs must more or less repeat the whole analysis after each reply by the opponent. Therefore the programs were probably not as handicapped by the two minute time limit in the experiment.

There is another explanation for why the experiment ranked some of the programs higher than humans of similar tournament strength. The scores in the test were based on the ability to find a correct move in individual, mutually *independent* positions, and not a correct *sequence* of moves in a whole game. A program may be able to find correct moves in a sequence of positions of the same game. However, although each of the moves may be correct, in a sequence they may not achieve a desired cumulative effect as they may belong to different plans, each of them winning alone but not if mixed with others. Therefore a program's individually correct moves may not in an actual game be as efficient as a human's sequence of moves, although possibly suboptimal consistently following the same plan.

Another weakness of the experiment may be that in some of the positions there is more than one good move. Our measures S, L and T were based on the comparison of *one* correct move with the move(s) proposed by the subjects, and therefore cannot be considered as absolutely reliable. One way of excluding this effect would be to base the interpretation of the results on the mutual similarity of subject's responses instead of the absolute correctness criterion. Subjects' responses would thus not be matched against correct responses in order to obtain the subject's success/failure pattern along the axis of 24 test positions. Instead, in order to find a similarity measure between two players, their responses would be compared directly, before matching them against the correct responses.

ACKNOWLEDGEMENTS

We would like to thank Professor Donald Michie for encouragement and helpful discussions of this work, our colleague Alan Shapiro for programming assistance, Morag Mullay for graphics work and Don Beal for useful comments. Also deserving many thanks are the various people who took the time and trouble to carry out our experiment on their computer chess programs and the human chess-player subjects who participated.

REFERENCES

Berliner, H. (1973) Some necessary conditions for a master chess program. *Proceedings of the Third International Joint Conference on Artificial Intelligence*, pp. 77-85.

Binet, A. (1894) *Psychologie des grands calculateurs et des joueurs d'echecs.* Hachette, Paris.

Bratko, I., Kopec, D. and Michie, D. (1978) Pattern-based representation of chess end-game knowledge. *Computer Journal*, 21, 2, 149-153.

Bratko, I., Tancig, P. and Tancig, S. (1976) Some new aspects of chess board reconstruction experiments. *3rd European Meeting on Cyb. and Sys. Res.*, Vienna.

Chase, W.G. and Simon, H.A. (1973) Perception in chess. *Cog. Psych.* 4, 55-81.

Elo, A. (1978) *The Rating of Chessplayers — Past and Present.* Batsford, London.

Groot, A. de (1965) *Thought and Choice in Chess* (edited by G.W. Baylor).

Mouton, The Hague and Paris. (Translation, with additions, of Dutch version of 1946.)

Kopec, D. (1977) Recent developments in computer chess. *Firbush News 7*, (edited by J.E. Michie). Edinburgh: Machine Intelligence Research Unit, University of Edinburgh.

Kmoch, H. (1959) *Pawn Power in Chess*. David McKay, New York.

Hort, V. and Jansa, V. (1980) *The Best Move*. RHM Press, New York. (Translation, with additions, of original Russian version of 1976.)

Matanovic, A. (1975) *Informator, No. 18*. Belgrade.

Michie, D. (1973) The path to championship chess by computer. *Computers and Automation*, Jan. 1973, 7-9, 36.

Michie, D. (1980) Chess with computers. *Interdisciplinary Science Reviews*, 5, 3, 215-227.

Michie, D. (1980a) Expert systems. *Computer Journal*, 23, 4, 369-376.

Nievergelt, J. (1977) Information content of chess positions: implications for chess-specific knowledge of chessplayers. *SIGART Newsl.* 62, 13-15.

Pachman, L. (1973) *Modern Chess Tactics*. Routledge & Kegan Paul, London. (Translated by P.H. Clarke from original Czech version of 1970.)

Philidor, A. (1749) *L'Analyse*. Paris.

Pitrat, J. (1980) The behavior of a chess combination program using plans. In *Advances in Computer Chess 2* (edited by M.R.B. Clarke), pp. 110-121. Edinburgh University Press, Edinburgh.

Shannon, C. (1959) Programming a computer for playing chess. *Philos. Mag. 7th Ser.*, 41, 256-275.

Simon, H.A. and Gilmartin, K. (1973) A simulation of memory for chess positions. *Cogn. Psychol.* 5, 29-46.

Slater, E.T.O. (1950) Statistics for the chess computer and the factor of mobility. In *Proceedings of the symposium on Information Theory*, pp. 150-152. Ministry of Supply, London.

Tan, S.T. (1977) Describing pawn structures. In *Advances in Computer Chess 1* (edited by M.R.B. Clarke), pp. 74-88. Edinburgh University Press, Edinburgh.

Turing, A.M. (1953) Digital computers applied to games. In *Faster Than Thought* (edited by B.V. Bowden), pp. 286-310. Pitman, London.