16 The Bratko-Kopec Test Revisited

T.A. Marsland

16.1 Introduction

The twenty-four positions of the Bratko-Kopec test (Kopec and Bratko 1982) represent one of several attempts to quantify the playing strength of chess computers and human subjects. Although one may disagree with the choice of test set, question its adequacy and completeness and so on, the fact remains that the designers of computer-chess programs still do not have an acceptable means of estimating the performance of chess programs, without resorting to time-consuming and expensive "matches" against other subjects. Clearly there is considerable scope for improvement, as the success of test sets in related areas like pattern recognition attest.

Here the performance of some contemporary chess programs is compared with earlier results from 1981, to help identify the properties of those cases that computers cannot handle well by search alone and to show the relative progress that has been made. Even though use of standard tests is still not widespread, many chess programming groups built such sets and a few have been circulated. One of the earliest was the NY1924 data set (Marsland and Rushton 1973) of about 800 positions, later used in a minor way to assess the performance of *Tech* (Gillogly 1978), and to develop evaluation function weighting factors (Marsland 1985). At about the same time Ken Thompson was building far larger test suites (Thompson 1979) and more recently Dap Hartmann worked with some 63,000 positions to extract knowledge from Grandmaster games (Hartmann 1987a,b). The Hartmann suite was used to tune the evaluation parameters of such programs as *Phoenix* and *Deep Thought*. When one considers that even 63,000 positions is a minuscule fraction of the estimated 10^{43} unique chess positions, what role can the small set of 24 B-K (Bratko-Kopec) positions play? Aside from being too small, the positions can be criticized because they consider only tactical and pawn lever moves, with many other important ideas and structures not covered. The tactical moves are now thought to be simple for computers, and also much larger test sets exist (Reinfeld 1945). Nevertheless, the true importance of pawn moves for high calibre play is brought out by the B-K positions better than by any other test set.

This chapter is an expanded version of a paper in the *New Directions in Game-Tree Search Workshop* proceedings, T.A. Marsland (ed.), Edmonton, May 1989, pp. 135-139. Later it appeared in the *Journal of the International Computer Chess Association*, vol. 13, no. 1, pp. 15-19.

Recognizing the narrow scope of the B-K suite, Jens Nielsen is developing a more sophisticated test with a greater range of features and is using it to estimate the Elo rating of commercial chess computers. Nielsen's (1989) system has many facets, using not only time taken to help measure a program's merit, but also testing the program's ability to reject moves. His system includes tests of endgame play, positional play, tactics and traps. At present some 145 problems are posed from 80 positions (many positions require the generation of a sequence of moves). Even though the test is time-consuming to apply, more than 40 programs have been tested and their Elo rating estimated with remarkable correlation to other accepted measures (Nielsen 1989). Like the B-K test and others, this system is of considerable benefit in the development of new chess programs, since it probes for the presence of specific knowledge and for the absence of common conceptual errors.

16.2 Previous Results

The original paper by Kopec and Bratko (1982) was also criticized for its unrealistic requirement that the program produce an ordered list of up to three choice moves. Although ordering moves is easy for humans, the pruning algorithm in most chess programs precludes consistent generation of such a list. That objection could have been overcome easily had the experiment been run slightly differently: by providing an ordered list of choice moves and rating performance according to the relative strength of the principal move proposed.

The last and final complaint aimed at prepared test sets is that programs can be tuned to perform well on the suite, perhaps at the expense of their overall playing strength. In principle, this objection is valid and serious, but in practice the pawn lever positions in particular have led to an appreciation of the importance of knowledge assessing critical pawn configurations. Also the harder tactical problems led to the development of selective search extensions (Anantharaman, Campbell and Hsu 1988) to identify and follow forced variations. Further, far more critical to the playing strength of programs than performance on any test suite are other factors, such as good use of time (Hyatt 1984; Anantharaman 1990), and effective use of transposition tables in the endgame (Nelson 1985). Nevertheless, it is clear from the results that the recognized best chess programs exhibit superior performance on the B-K test.

Consider Table 16.1 (Kopec and Bratko 1982), which shows an extract from the original results. Although the weakest programs fared badly when this test set was sprung upon them, some brute-force programs, notably *Belle*, *Duchess* and *BCP* did well even by today's standards. In particular, in 1981 *Belle* achieved a score of 18, which today is only exceeded by a handful of programs. Nevertheless, there can be no doubt that the comparably performing programs of today are stronger than *Belle*'81.

Turning now to the results of eight years later, Table 16.2 and Table 16.3, present the data supplied by by applicants to the 6th World Computer Chess

	Computer Subjects				
	Program	Rating	Score	T	L
1.	Chess Challenger ‘10’	Unr	1	1	0
2.	Chess Challenger ‘7’	Unr	5	2	3
3.	Sensory Chess Challenger	Unr	5	3	2
4.	Sargon 2.5	1720˜	5	2	3
5.	AWIT	1400	5	4	1
6.	OSTRICH81	1450˜	6	4	2
7.	CHAOS	1820	6	5	1
8.	Chess Champion Mk V (E)	1885˜	6.83	5	1.83
9.	Morphy Encore	1800˜	9.33	6	3.3
10.	BCP	1685˜	13	10	3
11.	DUCHESS	1850	16.50	10.5	6
12.	BELLE	2150	18.25	11	7.25

Key: (E) Experimental version; ˜ Rating is an estimate; (Unr) Unrated;
(T) Tactical score; (L) Score on pawn lever positions.

Note: Programs running off mainframe computers have names entirely in upper case letters. Others are stand-alone microcomputer programs.

Table 16.1: An extract from the original (1981) Bratko-Kopec results.

Championships, plus some 1986 data for *Awit*’83. Of the twelve tactical positions, Table 16.2, about half the programs can solve nearly all (thus equaling the *Belle*’81 score). Further, virtually all the programs can solve far more than half the tactical positions. As these results show, the harder problems are positions 10 and 22, which are presented in Figure 16.1. However, there was no pattern to explain why the eight programs which successfully solved 11 tactical problems could not solve them all, since their failures were uniformly distributed across five different problems (positions 7, 10, 16, 18 and 22). Also, there can be little doubt that these top programs could be "tuned" to solve all twelve B-K tactical problems, but at what cost to their average playing strength? Equally it would seem that problems 1, 12, 14, 15, 16, 19 and 21 are within reach of solution by all contemporary programs, given enough effort. So in some sense those positions are a measure of minimal acceptable strength.

For the lever positions shown in Table 16.3, however, few programs can solve more than half, and only three positions can be solved by almost all the programs. In particular, problems 4, 6 and 8 seem easy enough for those programs that have the right knowledge. Interestingly, 13 of the 22 programs solved all three problems and the others only failed to solve one each! On the other hand, almost no program can solve the three most difficult (namely positions 2, 9 and 23), all of which involve a pawn sacrifice for positional gain, either specifically, or as part of the analysis of the principal variations. Figure 16.2 shows two representative positions. Not only are these problems difficult, but also it is possible that the few programs which were successful in solving them may just have been lucky. Even so, there are possibilities for improvement, since although 15 programs solved neither problem 9 nor 20, *Mephisto* was able to solve both! This suggests that *Mephisto* might contain special pawn knowledge not found in other programs.

16.3 Conclusion

Our data leads to the final questions. Is the B-K test good enough for estimating the performance of chess programs? Clearly not, since the suite is too small and not wide-ranging enough. Despite that shortcoming, are there still things for programmers to learn from the B-K test? Clearly yes, especially for new programs and those programs which are alone in failing a particular problem. Conversely, when several programs solve one problem, some programming error or lack of knowledge is preventing correct solution by the others. Finally, although more and more chess programs are incorporating selective extensions and dynamic width control in the deeper portions of the search, the results show that at least one fully selective search program, *Awit*'83, achieved a respectable score on the test suite even though it was selective at every level in its search, and even though in over-the-board play it had a checkered career. This suggests that in the middlegame one can do quite well with selective search, but in the endgame totally different knowledge, time control and more dynamic search depth limits are required. Lack of these features accounted for *Awit*'s relatively poorer endgame play.

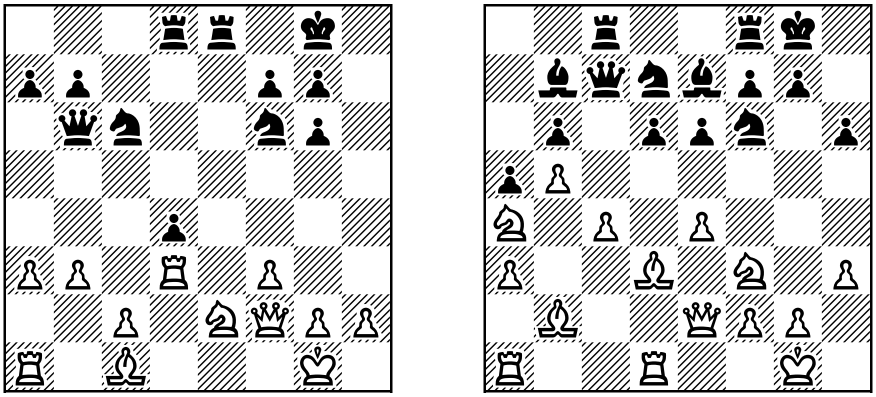
To conclude, the data presented here provides an opportunity to consider whether the calibre of a chess program is measured not so much by how many correct moves it makes in any test suite, but rather by the quality of the moves it proposes as alternatives to the acknowledged best choices. That is, the quality of a chess program is measured not so much by the frequency with which it plays optimal moves, but by the strength of its less than perfect choices.

Position Tactical (T)	1 Qd1	5 Nd5	7 Nf6	10 Ne5	12 Bf5	14 Qd2	15 Qxg7	16 Ne4	18 Nb3	19 Rxe4	21 Nh6	22 Bxe4	Ttl 12
AI Chess	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	12
Awit'83	ok	ok	Bd6	Qc5	ok	ok	ok	ok	ok	c5	ok	ok	9
Bebe	ok	ok	ok	Rd7	ok	ok	ok	ok	ok	ok	ok	ok	11
BP	ok	ok	Rg3	Qc5	ok	ok	ok	ok	ok	ok	ok	Nh5	9
Centaur	ok	e5	ok	Qc5	ok	ok	ok	ok	ok	ok	ok	e5	9
Cray Blitz	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	12
Dappet	ok	Bf4	ok	Qc5	ok	ok	ok	ok	f5	ok	ok	e5	8
Deep Thought	ok	ok	ok	ok	ok	ok	ok	Qh5	ok	ok	ok	ok	11
Hitech	ok	ok	Ra2	ok	ok	ok	ok	ok	ok	ok	ok	ok	11
Lachex	ok	ok	ok	ok	ok	ok	ok	ok	f5	ok	ok	ok	11
Mach 4	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	Ne5	11
Mephisto	ok	ok	Qc1	ok	ok	ok	ok	ok	ok	ok	ok	ok	11
Merlin	ok	ok	ok	Qc5	ok	ok	ok	ok	Be6	ok	ok	ok	10
Modul	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	12
Much	ok	Bf4	ok	Qc7	ok	ok	ok	ok	Bg4	ok	ok	Rd8	8
Pandix	ok	Rad1	Rg3	Qc5	ok	ok	ok	ok	Qb6	ok	ok	e5	7
Phoenix	ok	ok	ok	ok	ok	ok	ok	ok	Qb6	ok	ok	ok	11
Rebel	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	ok	Ne5	11
Shess	ok	Rad1	Bb4	Qc5	ok	ok	ok	Be7	Bg4	ok	Qe3	e5	5
Waycool	ok	ok	Ra2	ok	ok	ok	ok	ok	ok	ok	ok	Nh5	10
Y!89	ok	ok	Bb4	ok	ok	ok	ok	ok	Qb6	ok	ok	e5	9
Zarkov	ok	ok	ok	Qc5	ok	ok	ok	ok	f5	ok	ok	Rd8	9

Table 16.2: Results for the B-K tactical positions.

Position	2	3	4	6	8	9	11	13	17	20	23	24	Ttl
Lever (L)	d5	f5	e6	g6	f5	f5	f4	b4	h5	g4	f6	f4	12
AI Chess	e5	ok	ok	ok	ok	Rel	ok	Rac1	h6	Kb1	Bf5	ok	6
Awit’83	Rb1	a5	ok	ok	ok	Re1	ok	ok	e6	Qh5	ok	ok	7
Bebe	Ke3	ok	ok	ok	Nc3	Rc1	ok	ok	ok	Kb1	Bf5	bxc5	6
BP	e5	Qd8	ok	Kg4	ok	Bb5	Rfb1	Rac1	ok	Nb5	Bf5	c5	3
Centaur	e5	Qc7	ok	c4	ok	Re1	Nf5	Rec1	Qc8	Nc5	Bf5	exf5	2
Cray Blitz	g5	ok	ok	ok	ok	Bd3	ok	ok	c6	ok	o-o	ok	8
Dappet	e5	ok	ok	ok	ok	e5	ok	ok	h6	Nb5	Bf5	ok	7
Deep Thought	Kf3	Qd8	ok	ok	ok	Re1	ok	ok	c6	a3	Bf5	ok	6
Hitech	f5	Bd8	ok	ok	ok	Rel	Nf5	ok	a5	ok	Bf5	exf5	5
Lachex	e3	Rg8	ok	ok	ok	Bd3	ok	ok	h6	Qh5	Bf5	ok	6
Mach 4	Kf3	Rd8	ok	ok	ok	Rel	Nf5	ok	c6	Kb1	Bf5	exf5	4
Mephisto	Kf3	Bd8	ok	ok	ok	ok	Nf5	ok	c5	ok	Bf5	ok	7
Merlin	Kf3	ok	Nf3	ok	ok	g3	Nf5	ok	ok	Nb5	Bf5	ok	6
Modul	Kf3	Bd8	ok	ok	f6	Bb5	Rb1	ok	c5	ok	Bf5	ok	5
Much	e5	Rd8	ok	Kf3	ok	g3	Qa2	Rac1	Nb8	Nb5	Bf5	bxc5	2
Pandix	Kf3	Qd8	ok	ok	ok	Rel	ok	ok	c6	Qb5	Bf5	ok	6
Phoenix	Kf3	ok	ok	Kg4	ok	Rel	ok	ok	c6	Qh5	Bf5	ok	6
Rebel	Kf3	Bd8	ok	ok	ok	Re1	ok	ok	h6	ok	Bf5	ok	7
Shess	e5	ok	ok	ok	h4	Rel	ok	ok	b6	Nb5	Be6	bxc5	5
Waycool	f5	ok	ok	ok	ok	ok	Rfb1	Rac1	b6	Qh5	Bf5	f5	5
Y!89	e5	ok	ok	a4	ok	Bb5	Rfb1	Qe2	h6	Nb5	Bf5	exf5	3
Zarkov	e5	ok	ok	ok	ok	ok	ok	b3	h6	h3	Bf5	exf5	6

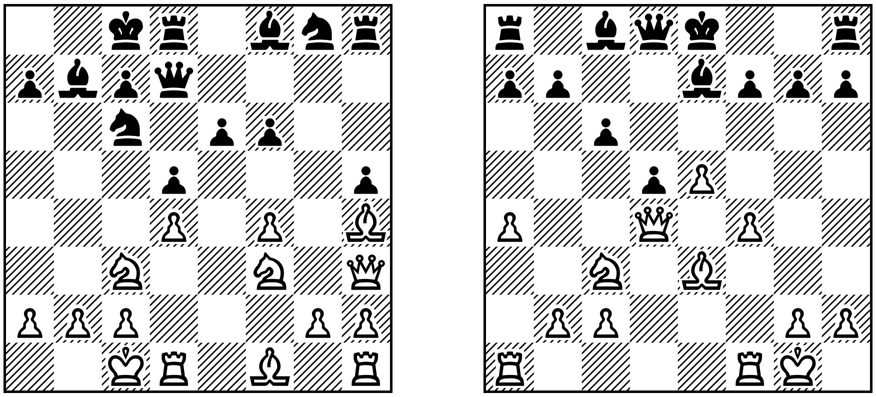
Table 16.3: Results for the B-K lever positions.



Posn. 10, black plays Ne5.

Posn. 22, black plays Bxe4.

Figure 16.1: Two difficult tactical positions.



Posn. 9, white plays f5.

Posn. 23, black plays f6.

Figure 16.2: Two difficult pawn lever positions.