NOTE

THE BRATKO-KOPEC TEST RECALIBRATED'

Shawn Benn and Danny Kopec

Department of Computer Science School of Computer Science

University of Maine, Orono Carleton University, Ottawa

Background and Purpose

As is well known, the B-K test due to Kopec and Bratko (1982) has, for a decade, been the only systematic and published means of assessment of chess-playing programs. This is not to detract from the practical value of Reinfeld's (1945) 300 positions; the difference is that the latter author's positions were not collected with a view of being tested by computers.

The B-K test (henceforth BKT) has enjoyed wide application and, as is inevitable with the tool of such long standing, has come in for a great deal of criticism as well. Critics have re that it was based on a very limited nutnber of positions (only 24). Proceeding from this fact it has been argued that, first, such a limited number could not provide a representative sample of the hugh set of all positions and, second that the selection of positions had been made in a manner ill-designed to be representative because it had concentrated, intentionally, on configurations bringing out only certain properties of the programs under test. We also note that the scoring in the BKT relies on the notion of best move; this notion may be stnbiguous in sotne positions. The present short note will argue that, in spite of its great age, the BKT still hits definite vsbdily.

Our experimental approach has been to administer lhe test to six selected chess engines on a testing platfonn as unifonn as possible. The selection has been governed by the availability of the machines and the results published below are riot claimed to have any absolute validity, which we specifically disclaim. Rather the machines have served as a test vehicle for the practical investigation of the BKT against present day, convnercially available chess engines.

The testees and the scoring

The identity and the main characteristics of the six testees are summarized in Table 1. The score for each testee was assigned as follows. The position was entered and the testee's suggested move was noted after 30, 60, 90 and 120 seconds. The testee was awarded 1 point when the move after 120 seconds coincided with the (known) best move. A fractional score of 1/4, 1/3, or 1/2 point was awarded when the (known) best move was indicated at 30, 60 or 90 seconds but not at 120 seconds. That is, the fractional scores corresponded to the maximum time at which the known best move was indicated as the program's choice.

Thus a testee is credited for having considered the best move even if that was abandoned later and moreover credited in rough proportion to the time for which he maintained that best move as its choice.

The raw scoring thus obtained served as an argument for a took up in Table 2, here reproduced from Kopec, Newborn and Yu (1986) in order to estimate each tester's rating.

---

1. This is major revision of the paper delivered under the title Comparison and Testing of Six Commercial Computer-Chess Programs on November 27. 1992 to the Workshop The Impact of Computer Chess on AI Research, Madrid, Spain [reported in the December 1992 issue of the ICCA Journal. pp. 228-2291. This event was held in conjunction with the 7th World Computer-Chess Championship, played there.
2. Presently affiliated to the Department of Computer Science, US Coast Guard Academy, New London.

Name	Author(s)	Hardware	Date of Release
CM3000	Software Toolworks, Inc.	80386 33 MHz	1991
Excel 68000	Fidelity Int., Miami, Florida	68000 12 MHz	1987
M_Chess	Martin Hirsch	80386 33 MHz	1991
Sargon IV	Dan and Kathe Spracklen, Spinnaker Software of Cambridge, Massachusetts	MacIntosh	1988
Sargon V	Dan and Kathe Spracklen, Activision of Palo Alto, California	80386 33 MHz	1991
Zarkov 2.5	John Stanback, Chess Laboratories of Pasadena, California	80386 33 MHz	1991

Table 1: The testees.

Score	Rating
0-4	1300- 1599
5-6	1600- 1799
7-8	1800- 1999
9-12	2000-2199
13-16	2200-2399
17-24	2400+

Table 2: Score on BKT vs. rating assigned.

Ever since its first publication, the BKT set was divided into two subsets of 12 positions each. Those conventionally marked T were designed to probe the testee's proficiency on

tactical positions, while those marked L were intended to estimate its capacity for finding a solution to strategical (lever) problems. Let S be the total score, T and L the part scores on the T and L positions in the order named. Then B defined as

$$B = 12x\ (T-L)/S,$$

may be interpreted as standing for the testee's bias towards tactical positions rather then strategical ones. Note that a negative value of B similarly indicates a relatively greater aptitude for strategic positions.

Results

From the results here given in Table 3, we present the following deductions.

Program	S	T	L	RatingR by Table 2	B
M.Chess	15	12	3	~2300	7.20
Sargon V	14	10	4	~2250	5.14
Zarkov 2.5	13	10	3	~2200	6.46
CM3000	13	8	S	~2200	2.77
Excel 68000	12.25	8	4.25	~2150	3.67
Sargon IV	11	7	4	~2100	3.27

Table 3: Scores (S., T, L) and the derived quantities (R and B).

- The general level of testees' play has gone up considerably since 1982 (Kopec and Bratko, 1982). There may be at least two distinct reasons for this fact, which moreover are not mutually exclusive:
  (1) The present testees are indeed superior to those tested a decade ago - not unlikely since the maximum time allowed was unchanged at two minutes and the hardware has speeded-up considerably since.
  (2) The engines' programmers have trained their programs on the BKT set and tuned them so as to maximize their scores.
- All testees are considerably biased towards tactical play as shown by their strongly positive. This accords well with the accepted wisdom that computer-chess programs are strong on tactics. One piece of evidence often cited in support of this impression is many programs' preferred move ordering, with capture moves at the top of the move list. This induces an eagerness for capture, which, in turn, may be interpreted as a preference for tactical play.
- Referring to Table 4 where we confront our estimated ratings (colum 5 of Table 3) with the ratings from the Swedish rating list where available for our testees we

remark that while absolute ratings show considerable discrepancy, the order of the ratings derived from the BKT and Swedish inter machine play is the same.

Program	BKT rating	Swedish rating
M_Chess	~ 2300	2127
SargonV	~ 2250	n.a.
Zarkov 2.5	~ 2200	2018
CM3000	~ 2200	1938
Excel 6800	~ 2150	1915
Sargon IV	~ 2100	n.a.

Table 4: BKT ratings compared with those of the Swedish Rating List.

Conclusions

Our experiment confirms the original experiment with BKT to the extent it indicates greater strength on the tactical position than on the strategic ones. This apparently is a characteristic still with us.

That the absolute scores deviate considerably from those reported by Mars (1990) and from those in the Swedish Rating List is no great cause for concern. It is important to note that, most significantly, the order of the performance ratings is identical for our simple and short test and for the Swedish Rating List based on hundreds of games for each testee. This, in turn, tends to show that the BKT is still an applicable and useful tool for a "quick-and-dirty" estimate of a program's prowess. If greater conformity with other published ratings is desired, this is easily achieved, we suggest, by suitably adjusting the values in column 2 of Table 2. This amounts to a recalibration. The fact that such a recalibration is feasible and leads to acceptable results is a strong indication that the BKT is still, for all its simplicity, a valid tool for program assessment.

The authors look forward to research correlating their resuhs with those to be obtained from Nielsen's (1991) chess-computer test set with its 86 positions.

References

Kopec, D. and Braiko, I. (1982). The Bratko-Kopec experiment: a comparison of human and computer performance in Chess. Advances in Computer Chess 3 (ed. M.R.B. Clarke), pp. 57-72. Pergsmon Press, Oxford. ISBN 0-08-026898-6.

Kopec, D., Newborn, M. and Yu, W. (1986). Experiments in chess cognition. Advonces in Computer Chess 4 (ed. D.F. Beal), pp. 59-79. Pergamon Press, Oxford. ISBN 0-08-029753-3.

Marsland, T.A. (1990). The Bratko-Kopec test revisited. Computers, Chess, and Cognition (edt. TA. Mars and J. Schaeffer), pp. 217-223. Springer Verlag, New York. ISBN 0-387-97415-6

Nielsen, J.B. (1991). A chess-computer test set. ICC'A Journal, VoL 14, No. 1, pp. 33-37.

Reinfeld, F. (1945). *Win at Chess*. McKay, New York. Reprinted (1958), Dover, New York. ISBN 0-486-20438-3.

---

3. The authors are grateful to the Editors for allowing them pre-publication access to the Swedish Rating List in the September 1993 issue.