Memory & Cognition
1982, Vol. 10 (3), 237-242

Recall or evaluation of chess positions as determinants of chess skill

DENNIS H. HOLDING
University of Louisville, Louisville, Kentucky 40292

and

ROBERT I. REYNOLDS
Nasson College, Springvale, Maine 04083

Previous research has found that the ability to recall briefly presented chess positions varies with playing strength, except when random positions are used. The suggestion therefore arises that mastery consists of recognizing configurations that are associated with plausible moves. This approach is tested by comparing the memory scores and move-choice protocols of players in six skill categories, using random chess positions. Contrary to any strong form of recognition-association hypothesis, differences in chess skill are shown to persist although memory differences are abolished. It is further shown that the moves selected are not based on those few pieces that are remembered. Skill-related differences in the accuracy of positional evaluations also occur, but they are less marked than in earlier results. An alternative approach to chess skill seems appropriate, in which memory effects may function at the evaluation phase.

Information processing models of chess skill have mainly taken their departure from de Groot's (1965) finding that players of different strengths were essentially alike in the number of moves they considered, in the depth of their search for move sequences and in other, similar measures obtained from spoken protocols. The only respect in which any substantial difference occurred was in their immediate grasp of new positions, as evidenced by recall after brief presentations. De Groot found that chess masters could reconstruct meaningful positions shown for between 5 and 10 sec with near-perfect accuracy, whereas weaker players showed far poorer recall.

The findings on brief recall have been repeated and extended several times (Charness, 1976; Chase & Simon, 1973b; Frey & Adesman, 1976; Goldin, 1978b, 1979) and represent the major experimental theme in the interpretation of chess skill. Chase and Simon (1973b) demonstrated clustering in the latencies for placement of pieces during the reconstruction of chess positions, finding that the pieces were grouped, or chunked, in terms of their degree of chess relationship. They suggest that recall is based on the retention of the labels for meaningful chunks in short-term memory and that masters have access to more and larger chunks than do weaker players. These differences disappear when randomized positions are used, presumably because predefined chunks are no longer available. Subsequent work indicates that briefly presented chess information may go directly into long-term store, since various processing activities interpolated between presentation and recall (Charness, 1976) or the presentation of a second position for subsequent recall (Frey & Adesman, 1976) have little effect on recall scores. Chess memory is also dependent on depth-of-processing variables, such that the activity of choosing a move, as against more superficial activities, assists later recognition (Goldin, 1978a) and assists recall as much as does intentional learning (Lane & Robertson, 1979). In each case, however, there is some correlation between memory scores and chess ability.

The association of ability with brief recall has been used by Chase and Simon (1973a) as a central feature in their explanation of skilled performance at chess, on the assumption that masters have access to a very large vocabulary of recognizable chess patterns. The number may be as large as 50,000, according to the Simon and Gilmartin (1973) model, which simulates the recall performance of stronger or weaker players by altering the number of piece configurations in long-term memory. Newcomers to the game take many years to acquire these learned patterns, with corresponding lags in both skill and recall ability. The crucial assumption of the Chase and Simon (1973a) model is that good, plausible moves are directly associated with each recognizable configuration of pieces, so that forward search and evaluation by skilled players may begin from the most promising initial moves. Although Chase and Simon take into consideration several other aspects of move generation and search, the efficiency of these information processes is governed by transactions involving the long-term storage capacities that vary with playing ability. Thus, what may be termed the recognition-association model sees the acquisition of chess skill as

0090-502X/82/030237-06$00.85/0

heavily dependent on building up recognition memory for a large number of familiar chess patterns.

As a result of this approach, there is a tendency to neglect processes other than memory. Simon and Chase (1973), for instance, consider that the structure of the search process will not differ greatly between masters and weaker players and that only the paths suggested by memorized patterns will be different. An emphasis on memory, at some expense to other components of chess skill, is already present in de Groot's (1965, p. 352) approval of the statement that "mastership consists of 'nothing but' experience and routine." The master's immediate perception of a new chess position is seen as differing from that of the novice, as a consequence of his readily accessible pattern memory. In a similar vein, Frey and Adesman (1976, p. 54) consider it established that "the ability to play well seems to depend on a learned perceptual skill rather than on the acquisition of a sophisticated problem-solving strategy."

Despite these conclusions, it has since been shown that players of different strengths do vary systematically in characteristics other than visual memory, in ways analogous to the differences between stronger and weaker computer chess programs. Holding (1979) has tested players whose U.S. Chess Federation (USCF) ratings range from 1,000 to 2,000; the scale, described by Elo (1978), reflects wins and losses in tournament and match play. It appears that players of different ratings differ substantially in the efficiency with which they assign numerical values to positional evaluations, a key component in successful forward search. Judgments by the stronger players discriminate more sharply between different chess positions, more accurately reflect the advantage ratios of the winning side, and more closely correlate with the pattern of results shown by successful computer programs. Investigations by Reynolds (in press) further qualify the earlier memory results by showing that the recall superiority of masters applies only to piece groupings about a functional center. Reynolds (Note 1) has also demonstrated a difference in search heuristics that did not appear in the protocols of Wagner and Scurrah's (1971) subject. Masters show a tendency, which is greatly reduced in weaker players, toward narrowing their search after encountering a favorable evaluation while broadening the number of moves considered after an unfavorable sequence. It should also be noted that Charness (1981), who matched older and younger players for level of skill, found that differences remained in breadth of search and in related measures that implied differences in memory for piece movements.

The view that chess mastery is based primarily on memory for piece configurations may therefore be incorrect or incomplete. The matter may be clarified by showing whether skill differences in choosing moves will persist in circumstances in which differences in memory performance are abolished. Given a positive outcome, this procedure will rule out any strong form of recognition-association theory in which pattern memory, as conventionally estimated by the brief recall task, is held solely responsible for chess mastery. A weaker form of the theory, which might view pattern memory as one of several components of chess skill, will still remain tenable. With respect to the weaker form, the proposed experiment should make some contribution toward assessing the relative importance of the recognition component.

Equating the recall scores across different skill levels is easily achieved by using random chess positions, as exemplified in previous research. Scoring any residual differences in skill presents a less tractable problem, since there exist no completely objective criteria. A promising approach, however, is to subject each test position to exhaustive analysis, to select the best move sequences or alternatives, and then simply to score whether or not the moves suggested by players correspond with the ideal sequences. A subsidiary aim of the experiment is to collect further evidence on the process of positional evaluation. Since the process of evaluation may itself be affected by the memorability of chess positions, both an initial first-impression score and a later, considered score are collected. The experiment therefore consists of briefly exposing random positions and obtaining (1) attempts at reconstruction from memory, (2) first-impression evaluations of the corrected positions, (3) move generation from the correct positions, over an extended period, and (4) final evaluations derived from the move analysis.

METHOD

Subjects

Chess players with known skill ratings were recruited individually from local clubs and tournament sites. There were 24 players, all male, 4 in each of the six USCF rating classes from Category V through Expert. The mean ratings in each category were, rounded to whole numbers, 1,065, 1,317, 1,501, 1,709, 1,914, and 2,129.

Materials

The one practice position and three test positions were each presented on a standard tournament-size chessboard with Staunton pieces. Brief exposures for recall were achieved by manipulating a board-sized lid and timed by a center-sweep second hand; a chess clock was employed for the choice-of-move procedure.

The stereotyped position developed by de Groot (1966, p. 42) on the basis of statistical likelihoods after the 20th move was used as the practice position while familiarizing players with the procedure. The three test positions were arrangements of 24 pieces (with one bishop, one knight, and two pawns removed from each side), corresponding with the state of the average game between Moves 20 and 25, at which most exchanges and gains are made (Holding, 1980). The set of positions was assembled by using random-number tables to select pieces and assign their placements until positions were reached that satisfied the constraints that (1) neither king was in check, (2) no pawn occupied the first or eighth ranks, and (3) no piece was attacked by pawns or otherwise attacked while undefended.

Recorded in Forsyth notation (commas conclude ranks, from the eighth down; lowercase = black, uppercase = white; numerals represent spaces), the positions were: (1) n1B2b2, r1p3p1, P3Rp2, 2PqP1k1, 5pP1, 2Rp1p1K, 1P1P1r2, 2Q3N1; (2) 8, 1p2R2n, 1k1pPr2, 3P1ppq, 3P1p1p, PBKR3Q, 2PP4, 1r4bN; and (3) n7, p5R1, 1qp1PQ2, pk1pP3, r4NpP, B1P1p2b, 1PK1P2r, 6R1.

The evaluation scale followed Holding (1979) in calling for positional judgments such that, given a score of 10 for the weaker side, the stronger side is awarded a value from 10 through 20 points. The scale is provided with supplementary verbal anchors (for example, 14 is said to represent a "clear advantage"), but consistency in using the numerical values is stressed.

Procedure

Each player perused a sheet of instructions and a copy of the evaluation scale. He then performed a "dress rehearsal," using the practice position, followed by the same procedure on all three test positions. In each case, the board was exposed for 8 sec, the pieces were removed to the side, and he attempted to reconstruct the position, with the requirement that all pieces be placed on the board. The experimenter then recorded the reconstruction, corrected the position, and asked for an immediate first impression of the evaluation score. Following this, the chess clock was started and the player was given 3 min in which to deliberate on the best move (for white, in all cases), while studying the corrected position. When the flag fell, he announced his decision, the other side of the clock was started and he received 2 min in which to suggest an explicit sequence of further "best" moves by black and white, together with any comments on the position. Each sequence consisted of a single line of play, rather than the generation of a move tree, to ensure objectivity in scoring. Finally, the player was asked for a considered evaluation score.

Each of the 12 possible orders of presentation of the 3 positions was used twice, randomly assigned to players. Half of the players saw each position as white, and half as black, in each playing class.

RESULTS

Memory and Choice of Move

The memory scores simply represent the number of pieces correctly placed, averaged over the three test games. The choice-of-move protocols were scored by comparison with extensive analyses made by the authors (R.I.R., who has held a master rating, is currently rated 2,100+; D.H.H. is currently rated 1,700+). The best initial moves for white and black appear to be: (1) Q-K1, answered by BxP, Q-Q5, or P-B4; (2) B-R2 and R-N4; and (3) Q-B5 and R-K5, or (transposable) P-K7 and N-B2. Subsequent optimal moves were ascertained through the fourth move for black in most variations. The scoring system was ad hoc in character, but it was designed to reflect the quality of the suggested move sequences. The basic score consisted of awarding one point for each listed move recorded by a player in the correct sequence. A sequence was not deemed interrupted by a forcing "zwischenzug" (in-between move), but after a neutral move only one further good move was scored, unless the transposition clearly led to no disadvantage. In addition, since some moves by weaker players were extremely bad, up to one move per game was given a minus score if leading directly to mate or to uncompensated loss of material. Finally, in order to recognize creativity without undue sacrifice of objectivity, up to one divergent good move per protocol might be credited if earned. Both authors first scored the protocols independently, any differences being subsequently resolved by discussion and further analysis.

Both the recall scores and good move scores are shown in Figure 1 in the form of means for each USCF rating category. As expected, the memory scores showed no significant correlation with the level of skill indicated by players' USCF ratings [$r(22) = .10$]. On the other hand, the good move scores did increase significantly with rating strength, as predicted [$r(22) = .75$, $p < .001$]. Although unpredicted, it is also of interest that the mean number of moves suggested in the protocols tended to increase with rating strength [$r(22) = .44$, $p < .01$]. The differences were small; the mean numbers of moves per position were 5.8 (for 1,000-1,399 players), 6.2 (for 1,400-1,799), and 7.6 (for 1,800+).

Specific Recall Analyses

It might be insufficient to show merely that the choice-of-move scores improve despite poor overall recall performance; it could be argued that the chosen moves are based on the few pieces that were selectively retained in the memory test, since these indicate the recognizable configurations that are associated with good, plausible moves. This is unlikely, since it seems to imply that the mean 6.54 moves (plies) suggested per position are derivable from the mean 6.15 pieces remembered. However, a further analysis of the protocols was undertaken in order to establish what proportion of the suggested moves was based on the specific pieces remembered by each player; the baseline number of suggested moves was diminished by considering only those using different pieces, discarding second and later moves by the same piece.

The players were considered in two skill groupings for initial analysis, since the hypothesized tendency

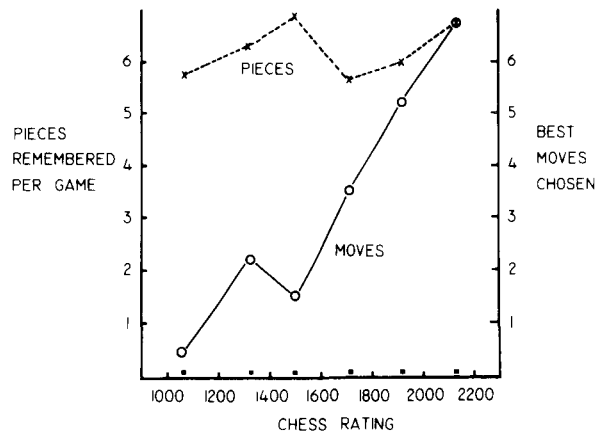

Figure 1. Memory scores compared with good move scores at each USCF rating category.

might depend on skill level. The stronger players (1,600 and above) had accurately remembered a mean of only .198 of the different pieces moved, whereas the weaker players (below 1,600) had similarly remembered .217. The difference does not approach significance [$t(23) = .19$], and both groups were combined for further analyses of the specific hypothesis that players might base more than a chance proportion of their moves on remembered pieces.

The numbers of remembered pieces that were moved, expressed as proportions of all pieces moved by each player, were therefore used in a further test. The proportions of remembered pieces that were moved should significantly exceed the proportion unmoved, if move choices are based on memory. The proportions moved were therefore compared with the proportions of all unmoved pieces that were nevertheless remembered. (Thus, $r \cap m/m$ was compared with $r \cap \bar{m}/\bar{m}$, where r = remembered and m = moved.) The difference between the two sets of proportions was significant [$t(23) = 2.48$, $p < .05$], but in the unexpected direction that the proportion of remembered pieces that players failed to move was larger than the proportion that they did move. Further analysis showed that this occurred because of a tendency toward differential recall and differential choice for moves of pawns vs. nonpawn pieces. A greater proportion of pawns was accurately remembered [$t(23) = 2.15$, $p < .05$], presumably because pawn chains form a useful anchor for recall. On the other hand, nonpawn pieces were more frequently used in move choices [$t(23) = 15.02$, $p < .001$], as happens in natural games.

As a final check on the hypothesis, the proportions were determined separately for those moves considered good according to the criteria above, since it could be argued that poor moves were those generated in the absence of remembered patterns. In fact, the proportion of good moves with remembered pieces appeared smaller than the proportion of poor moves with remembered pieces, but the number of good moves was relatively small and the difference was not significant [$t(23) = 1.87$].

Evaluation Scores

The positional evaluations were difficult to make because of the sheer complexity of the positions and the many tactical diversions, but were deemed by the authors, in advance of scoring, to be (1) 16 for white, (2) 11 for black, and (3) 16 for white. The "Chess 4.5" computer program, used as a comparison by Holding (1979), would wrongly have given (1) a small advantage (38.5) for black, (2) a solid advantage (81.9) for white, and (3) a strong advantage (117.6) for white. The observed evaluation scores, measured against the arbitrary standards, were highly variable both within and between game positions. Averaging across positions, neither the first nor the second of the overall mean evaluation scores correlated significantly with rating

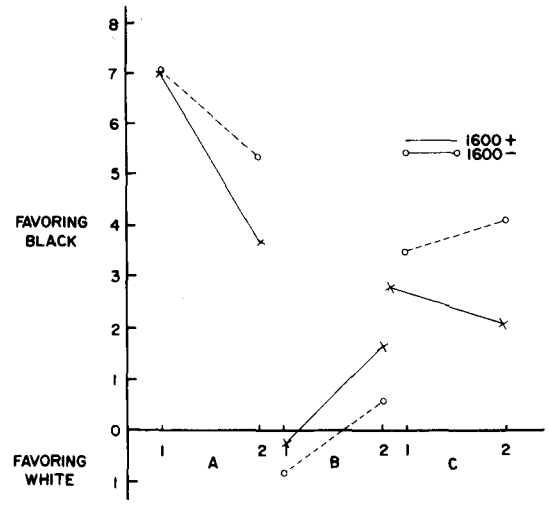

Figure 2. Changes in accuracy of evaluation by strong and weak players for each game position.

strength [$r(22) = .15$ and $r(22) = .09$, respectively]. However, the stronger and weaker players differed in the way their evaluations of the three positions changed. Figure 2 shows the accuracy of both sets of evaluations, measured in terms of algebraic deviation from the ideal scores proposed above.

In Position A, the evaluations of the stronger (1,600+) players changed significantly (sign test: $p < .05$) in the direction of greater accuracy, whereas a lesser change for the weaker players (1,600–) failed to approach significance ($p = .377$). In Position B, the evaluations were particularly difficult to score with confidence because black apparently had the best practical chances, although all variations could be drawn. The stronger players shifted significantly from favoring white toward favoring black ($p < .05$), whereas the weaker players apparently shifted less ($p = .969$); for the stronger players to be considered correct, the evaluation would have to be changed to 13 in favor of black. Position C shows the stronger players improving, at a marginal level of significance ($p = .055$), whereas the weaker players did not significantly worsen ($p = .377$). No further analyses of the evaluations showed significance.

DISCUSSION

The data leave no doubt that the quality of the moves chosen by variously skilled players is independent of the level of players' recall performance. Furthermore, the nature of the chosen moves is not connected with the identity of the few pieces that are remembered. These findings seem to argue strongly against de Groot's (1965) position and to encourage some revision of Chase and Simon's (1973a) interpretation of the relationship between playing strength and recall. The observed tendency toward better memory for pawns, contrasted with the preference for moving nonpawn pieces, seems to

exacerbate the difficulty. Thus, to the extent that long-term memory may be organized around pawn chains, it becomes less plausible to contend that skill depends upon immediate perception gaining access to pattern-move associations.

It is possible to question the conclusiveness of the present experiment on two related grounds, stemming from the use of a combination of brief exposures for recall together with a prolonged period of inspection for the choice of moves. Although the brief-exposure technique is the one customarily employed and the one that provides the empirical basis for the recognition-association model for memory and chess skill, it might be argued that the time available for move choice should have been as brief as the presentation time for recall; testing would thus be conducted in the conditions appropriate for speed play. This procedure would have somewhat lowered the general level of good move selection in these very difficult game positions, but it seems unlikely that it would affect the main outcome other than by loss of sensitivity. It is well known that differences in normal playing strength remain evident in high-speed play, so that many grandmaster games at speed chess are difficult to distinguish from tournament games (Church & Church, 1977), although these findings are derived from meaningful rather than from random positions. In any case, it is considered play of the tournament kind that one wishes to explicate.

A second objection concerns how much is remembered when an extended time is given for inspecting the position. The Chase and Simon (1973a) recognition-association model derives its support from the findings on recall scores but implies a recognition process for identifying chess patterns that might continue to operate during the choice-of-move period. It is certainly possible that more configurations are later recognized than are originally recalled; in fact, Goldin's (1979) results indicate that highly skilled players achieve good recognition scores with long exposures to scrambled positions, although yes-no decisions on whole-board positions may preserve rather little detailed information. On the other hand, if the period for recognition is extended in this way, the model no longer accounts for de Groot's (1965) contention that chess masters immediately perceive a novel position in a different way, nor for the fact that his grandmaster subjects considered initial moves that never appeared in expert protocols. In any case, there are many other processing stages in chess-move analysis in which memory effects might be expected to impinge. It therefore seems premature to postulate that the part played by memory is confined to the association of plausible moves with recognized subpatterns.

An alternative, computer-based approach to move selection would assume that, in addition to generating a set of initial moves, it is necessary to project ahead a branching tree of replies and countermoves (until a quiescent position is reached, or until time and capacity are exhausted), to assign values to the resulting end positions, and to back these values up the tree by a form of minimax procedure such that the initial move with the most favorable outcome is selected for play. A good deal is known concerning these stages in human move selection. Reynolds' (in press) data show that masters and weaker players direct their initial attention to different areas of the board; they also differ in the specific search heuristics mentioned above. It is also considered by many authors (cf. Hearst, 1977) that human players differ from computers in directing their searches toward long-range goals. Then, despite de Groot's (1965) result, it appears that differences do exist in the depth of search by stronger and weaker players (Charness, 1981; Wagner & Scurrah, 1971).

Differences at the evaluation stage are among the most critical to computer programs and may be demonstrated in human players by further analysis of Holding's (1979) data. A gross count of those evaluations that give the wrong side as winning, with half-errors for assigning draws in winning positions, shows a significant difference between players in different USCF rating classes [$F(4,45) = 2.98$, $p < .05$]. Similar differences in the accuracy of evaluations appear in the present experiment, despite the difficulties presented by the random positions, with stronger players improving significantly in two of the test positions. These conclusions should be qualified by the finding that human evaluations show a move-oriented bias; move selection and positional evaluation tend to interact recursively, in a manner reminiscent of de Groot's (1965) "progressive deepening." If the evaluation made is partially dependent on the move chosen, the cases in which players choose the correct move when the winning side is to move (or the wrong move when the losing side has the move) should give higher evaluation scores than in the opposite cases. For the stronger players in Holding's (1979) study, who make a sufficient number of correct moves for analysis, the predicted difference does arise significantly (Wilcoxon $T = 88$, $p < .05$].

All of the stages that have been mentioned might be influenced by the accumulated knowledge of previous chess experience. However, the evaluation stage is particularly crucial to skill at chess, since projecting ahead is wasted if one does not know whether or not a potential end position is favorable. Efficiency in evaluation seems to be difficult to acquire, leading to consistent differences between differently skilled players that might plausibly be ascribed to differences in memory for positional information. The present evidence is not conclusive, since, with the use of random positions, efficiency in evaluation appears to have suffered some deterioration, although some skill differences persist. Problems in evaluation are perhaps as likely to result from the extreme difficulty of the positions as from the absence of remembered patterns. However, it would not be unreasonable to assume that it is the dependence of evaluation upon memory in some form that underlies

the commonly observed correlation between memory and chess skill and, hence, that the decline in evaluation efficiency is a probable consequence of the poor memory scores for random positions.

The hypothesis that memory determines evaluation is consistent with the finding that no skill differences appeared in the first-impression score. At the same time, the fact that the stronger players improved their evaluations between the first and second attempts, in at least two of the three positions, suggests that a substantial component of their memory organization takes a form that is not readily available to direct recognition. It should be noted that transferring the influence of memory from the move generation to the evaluation stage need not imply a theory in which the effects of a large repertory of specific chess patterns are utilized. Memory effects at the stage of positional evaluation might consist of applying general principles, perhaps as an elaborated set of common precepts (rooks belong on open files; backward pawns provide targets; etc.) or, instead or additionally, might represent the direct recognition of previously occurring favorable positions. In either case, ascribing the relation between memory and chess skill to the stage of evaluation appears to offer a viable alternative to explanations in terms of a direct recognition-association model.

REFERENCE NOTE

1. Reynolds, R. I. *Search heuristics of chessplayers of different calibres.* Paper presented at the annual meeting of the Southern Society for Philosophy and Psychology, Louisville, April 1981.

REFERENCES

CHARNESS, N. Memory for chess positions: Resistance to interference. *Journal of Experimental Psychology: Human Learning and Memory,* 1976, **2**, 641-653.

CHARNESS, N. Search in chess: Age and skill differences. *Journal of Experimental Psychology: Human Perception and Performance,* 1981, **7**, 467-476.

CHASE, W. G., & SIMON, H. A. The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing.* New York: Academic Press, 1973. (a)

CHASE, W. G., & SIMON, H. A. Perception in chess. *Cognitive Psychology,* 1973, **4**, 55-81. (b)

CHURCH, R. M., & CHURCH, K. W. Plans, goals, and search strategies for the selection of a move in chess. In P. W. Frey (Ed.), *Chess skill in man and machine.* New York: Springer-Verlag, 1977.

DE GROOT, A. D. *Thought and choice in chess.* The Hague: Mouton, 1965.

DE GROOT, A. D. Perception and memory versus thought: Some old ideas and recent findings. In B. Kleinmuntz (Ed.), *Problem solving: Research, method and theory.* New York: Wiley, 1966.

ELO, A. E. *The rating of chessplayers, past and present.* New York: Arco, 1978.

FREY, P. W., & ADESMAN, P. Recall memory for visually presented positions. *Memory & Cognition,* 1976, **4**, 541-547.

GOLDIN, S. E. Effects of orienting tasks on recognition of chess positions. *American Journal of Psychology,* 1978, **91**, 659-671. (a)

GOLDIN, S. E. Memory for the ordinary: Typicality effects in chess memory. *Journal of Experimental Psychology: Human Learning and Memory,* 1978, **104**, 605-611. (b)

GOLDIN, S. E. Recognition memory for chess positions: Some preliminary findings. *American Journal of Psychology,* 1979, **92**, 19-31.

HEARST, E. Man and machine: Chess achievements and chess thinking: In P. W. Frey (Ed.), *Chess skill in man and machine.* New York: Springer-Verlag, 1977.

HOLDING, D. H. The evaluation of chess positions. *Simulation and Games,* 1979, **10**, 207-221.

HOLDING, D. H. Captures and checks in chess: Statistics for programming and research. *Simulation and Games,* 1980, **11**, 197-204.

LANE, D. M., & ROBERTSON, L. The generality of the levels of processing hypothesis: An application to memory for chess positions. *Memory & Cognition,* 1979, **7**, 253-256.

REYNOLDS, R. I. Search heuristics of chess players of different calibres. *American Journal of Psychology,* in press.

SIMON, H. A., & CHASE, W. G. Skill in chess. *American Scientist,* 1973, **61**, 394-403.

SIMON, H. A., & GILMARTIN, K. A simulation of memory for chess positions. *Cognitive Psychology,* 1973, **5**, 29-46.

WAGNER, D. A., & SCURRAH, M. J. Some characteristics of human problem-solving in chess. *Cognitive Psychology,* 1971, **2**, 454-478.

(Received for publication October 27, 1981;
revision accepted February 18, 1982.)