ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 62, No. 1, April, pp. 63–69, 1995

# Characteristics of Skilled Option Generation in Chess

GARY KLEIN, STEVE WOLF, LAURA MILITELLO, AND CAROLINE ZSAMBOK

*Klein Associates Inc., Fairborn, Ohio*

**We tested the hypothesis that skilled decision makers generate satisfactory options as the first ones they consider, thereby avoiding the need to perform extensive generation and evaluation. A total of 16 chess players were divided into groups of high and medium skill ratings. Four different chess positions were presented, and the task of the subjects was to generate the next move to be played, articulating every move they thought of, no matter how weak. Only a small proportion of legal moves were worth playing, but the majority of initial moves that the subjects generated were good ones, with a mean rating of 3.59 on a 5-point scale. These results suggest that decision makers may be able to pursue deep rather than broad searches, relying on their ability to generate feasible options as the first ones considered.** © 1995 Academic Press, Inc.

## INTRODUCTION

Although option generation is a critical step in decision making and problem solving (Keller & Ho, 1988; Serfaty, Entin, & Riedel, 1991), it is not a well-understood phenomenon. Standard accounts of decision making rely on a filtering concept in which a large option set is successively refined until a superior option is identified (Janis & Mann, 1977) or a moderately sized option set is identified and evaluated to select one (Gettys, Pliske, Manning, & Casey, 1987). In contrast, naturalistic models of decision making, such as the Recognition-Primed Decision Making model (RPD) (Klein, 1989), dispense with a filtering concept and posit that people can recognize a situation as typical, thereby calling forth typical reactions without having to sift through large sets of alternatives. Therefore, we are comparing a two-stage versus a one-stage model. The RPD model describes a decision maker using serial generation and evaluation of options so that it is not necessary to go through a lengthy elimination or choice process. The model also assumes a satisficing criterion, so that it is only necessary to generate a single feasible option in order to accomplish a task. Support for the model comes from retrospective protocol analyses from more than 480 decisions in field settings such as firefighting, design engineering, and military command and control.

The purpose of this study was to build on our understanding of the option generation process by examining whether subjects are able to generate feasible options as the first ones considered. This prediction reflects the RPD model's emphasis on the role of experience in decision making. The prediction that experienced decision makers will be able to generate reasonable options as the first ones considered is not inconsistent with alternative decision models, hence this is not a strong test of the RPD model. Nevertheless, the alternative accounts have not suggested the possibility that people can use experience in this way.

A simple means by which we can determine whether subjects are generating reasonable options early in the sequence is to compare the quality of these options against the quality of a complete set of options for some finite problem space. If these early options are better than what might be expected by drawing randomly from the pool of all available options, then the hypothesis would be supported. We are not claiming that any decision researchers hypothesize random option generation. However, we believe it is important to supply the empirical evidence in support of this non-random claim.

de Groot (1946/1965) stated about chess players that, "Often one finds, abstracts, 'sees,' immediately from the structure of the situation on the board what is essentially going on and, therefore, what must be done" (p. 306). The data gathered from de Groot's protocols

The authors gratefully acknowledge the contribution of our chess Master, Tony Mantia, whose guidance and assistance were invaluable across all phases of the project. We thank each of the subjects from the Dayton Chess Club for their eager participation in the study. We also thank Adrian de Groot, Len Adelman, John Payne, and Charles Gettys for their insightful reviews of earlier drafts of this manuscript. These comments were extremely valuable in helping us to generate the present paper. Special thanks are extended to Dennis Holding for his assistance in the preliminary stages of the study, in addition to his comments on the manuscript. Finally, we acknowledge the U.S. Army Research Institute (Contract MDA903-89-C-0032) for their funding of this research project. Address reprint requests to Dr. Gary Klein, 582 E. Dayton-Yellow Springs Rd., Fairborn, OH 45324.

0749-5978/95 $6.00

have shown that highly experienced chess players evaluate generated moves through progressive deepening; the mental "gaming out" of potential moves, and discarding poor moves until an acceptable one is found. In reviewing de Groot's protocols for the first three boards we found that in most cases (about 73% of the time) highly skilled players generated just four or five options, and investigated only a subset of these. More important for the purposes of this study, skilled chess players in de Groot's study were able to generate strong moves as the initial ones they considered.

If options are randomly generated then, on average, the first option an individual formulates should be no better than the tenth, or the twentieth that comes to mind. If there is an ordered generation of options, then the initial ones should be the strongest, and the last ones should be the weakest. We posit that subjects will use a serial evaluation strategy thereby limiting the size of the set to a small number of reasonable options. Our method of testing this idea was to use chess positions from several different games. We presented a game position to subjects and asked them to name possible moves as they thought of them. All moves formulated by the subjects were recorded. If the initial moves were randomly generated, the quality should have been no better than that of any of the other possible moves the subject could have made. If the moves supplied by the subjects were significantly better than what might be expected by chance, then the hypothesis of nonrandom, ordered option generation would be supported.

It should be noted that we are not adopting the strong view of Chase and Simon (1973), who suggested that chess skill could be explained by direct recognitional matches to patterns stored in memory. Holding (1985) has forcefully argued against this position, and de Groot (1987) has also pointed out that simple recognitional processes are not sufficient for explaining the way good chess moves are constructed. If this hypothesis were true, we would expect to see subjects attempting to recall moves that they had used before. In the body of research that we have examined, there has been no evidence to support this position. We believe that the construction of chess options is based on pattern recognition but cannot be reduced to memory retrieval. We do not know how much of the option-generating process is contributed by memory retrieval and how much is constructed. It remains an open question.

A second issue in our study was the impact of skill level on the generation of options. A naturalistic model of decision making would posit that more experienced decision makers should not only select moves of higher quality but also generate options of higher quality. Therefore, we studied chess players at two different skill levels.

A final component of this study examined the effects of context on the quality of moves generated. We manipulated context to determine its effect on the quality of options generated. In one variation a game position was shown and the subject was asked to supply possible moves. In the second, the subject was shown a series of preceding moves from a particular game, giving the subject the opportunity to observe the "flow" of the game as it developed, before being asked to identify moves in the game position. If context impacts option generation, then the moves generated in this second condition should prove significantly better than those moves generated without context.

To summarize, our major hypothesis is that subjects will generate options in an ordered fashion (based on move quality), not randomly. Additional hypotheses are; first, that subjects will generate relatively small option sets; second, highly skilled subjects will generate better options than less skilled subjects; and third, that context will improve the quality of options generated by both highly and less skilled subjects.

## METHOD

### Subjects

Sixteen subjects (all males) participated in the study. All of the subjects were members of an area chess club who volunteered to participate. They were divided into two groups of skill level based on the United States Chess Federation (USCF) rating system. Eight individuals with a rating between 1150 and 1600 were placed in the medium skill level group. Another group of eight individuals with a rating of 1700–2150 were placed in the high skill level group. The average rating of the medium skill group was 1333, hereafter, Class C players. The average rating of the high skill group was 1884, hereafter, Class A players. While the typical USCF categorizations of Class A and Class C players do not include such a broad range of skill ratings, we adopted these labels for the purposes of our study. All subjects from each group participated in all phases of the experiment. Subjects were paid 10 dollars each for their participation in the experiment.

### Stimulus Materials

*Board preparation.* To test the stated hypotheses, four board configurations (hereafter, "Boards") from back issues of CHESS magazine were obtained. The boards represented either middle games ($n = 1$) or end games ($n = 3$). The labels "middle game" and "end game" are commonly accepted by chess players to designate the general degree of completeness of a game. Initially, we selected boards on the basis of presumed equivalence. However, we discovered that subjects reacted to the middle game board (hereafter, Board A) differently than they did the other three boards. Board equivalence was determined by a chess Master (rating = 2400), based on (1) relative material strength, (2) relative position strength, and (3) relative time strength. These same criteria were used within each board by the chess Master to determine equality of the opposing sides.

To prepare the boards for the context condition, the chess Master generated three moves (actually pairs of moves or plies) leading up to the current position. The number of prior moves to be shown ($n = 3$) was selected in an effort to balance the subjects' ability to understand the "flow" of the game while not allowing too much time for hypothesis generation to begin.

*Ratings of move quality.* To evaluate the quality of the possible moves for each of the boards, a 5-point rating scale was constructed with the assistance of our chess Master. The possible ratings that could be given a move were: 5—best possible move; 4—a good, strong move; 3—average move, better moves available; 2—poor move, many better moves available; and 1—a blunder. This scale was developed to allow the participants in the study to provide their subjective ratings of the moves. In addition, an objective source of move ratings was obtained for each of the four boards. These ratings were established by several groups of Grand Masters (size of the groups ranging from four to six members) who considered each of the boards and then discussed the relative merits of each of the promising moves. The Grand Master ratings, taken from CHESS magazine, were based on a point award system ranging from 10 points for the best possible move to zero points for moves unworthy of consideration.

### *Experimental Design*

The design was a 2 × 2 × 2 fixed complete factorial design with the following factors: skill level (high, medium); board (A, B, C, D); context (present, absent). We manipulated skill level as a between-subjects factor; the latter two were within-subjects factors. The dependent variables were the order of the generated moves, the number of moves generated, and the subjects' ratings of the moves.

### *Procedure and Materials*

Prior to the experiment, each subject was asked to bring his own chess board, pieces, and timer. This was done to take advantage of the subject's higher familiarity with his own equipment and thus reduce piece identification errors. There was one case in which a subject was unable to supply his own equipment. For this individual, materials were provided that were comparable to the subject's. Moves generated and move ratings were recorded using paper and pencil. In addition, the experimental sessions were audio taped as a means to resolve any discrepancies in the recording of moves generated. There were two occasions for which the audio recordings were used to clarify a generated move.

The experiment consisted of two sessions. Prior to beginning the first session, the experimenter described phase one of the experiment by reading a prepared text to a group of subjects (size ranging from 2 to 4). After the instructions were given, each of the subjects was led to a quiet area to begin the experiment. Subjects were told that in the first phase, they would be shown four boards, two with contextual moves leading up to the actual position and two without context. Then the subjects viewed the first board. At this time the experimenter indicated which side (white or black) the subject would be playing from and whether moves would be shown leading up to a position. If prior moves were to be shown, the experimenter announced the move, waited briefly (about one second), and then moved the piece. The experimenter repeated the process until the board was configured for the subject to generate the next move himself. At this point, the subject was asked to provide the experimenter with an ongoing verbal protocol of the moves he was considering, even those moves the subject would immediately reject. When the subject had selected a move from among the options being considered, he was to inform the experimenter so the next board position could be set up. The subjects were told that they would be given a 15-min time limit in which to select a move. Once all four of the boards had been tested the second phase of the experiment began.

In the second phase, a new set of instructions was read to the subject. In this phase, subjects were again shown each of the four boards. This time, however, the experimenter asked the subject to provide a rating, using the 5-point scale described above, for each of the legal moves from the current board position. No time limit was imposed on the subject to provide ratings. An average of approximately 30 legal moves was possible for each of the four boards yielding a total of 124 moves to be rated.

The subjects were given the opportunity to end the experiment at any time. None of the subjects exercised this option.

## RESULTS

### *H1: Random versus Ordered Option Generation*

*First move generated.* The primary hypothesis was that experienced chess players would generate feasible and acceptable moves as the first ones considered, rather than generating options randomly. This hypothesis was strongly supported by the data. Figure 1 shows the distribution of all possible legal moves ($n$ = 124), as rated by the subjects themselves, and also the distribution of the self-ratings of the first moves generated ($n$ = 64). Inspection of Fig. 1 reveals that the distribution of all legal moves is largely comprised of moves rated 1 and 2. The distribution for the actual first moves generated is clearly different. If subjects were randomly sampling from the legal moves, they would have shown a high number of moves rated 1 and 2, and a low number of moves rated 5. The actual distribution was just the reverse. A Chi-Square analysis of the observed and expected frequencies of these ratings was computed, using the ratio of all legal moves to generate the expected frequencies of quality ratings for the 64 first moves. The $\chi^2$ was significant, $\chi^2$ = 211.145, $p < .0001$, which demonstrates that the subjects were not randomly sampling from the pool of available moves. Furthermore, only 16 of the 64 first moves were given a rating of less than 3, showing that the subjects were fairly satisfied with the initial moves they considered.

We also examined the objective quality of the moves generated, using the Grand Master ratings. For this analysis, we designated all moves receiving any Grand Master points as acceptable ($n$ = 20), and moves receiving no Grand Master points as unacceptable ($n$ = 104). A total of 41 of the 64 first moves qualified as acceptable. Using the Grand Master ratings for all the legal moves as the criterion, we generated proportions for the expected frequencies of moves with and without Grand Master points, and compared these to the observed frequencies (see Table 1). The result, $\chi^2$ = 113.9, $p < .0001$, again suggests that the subjects were not randomly sampling from the pool of all possible options.

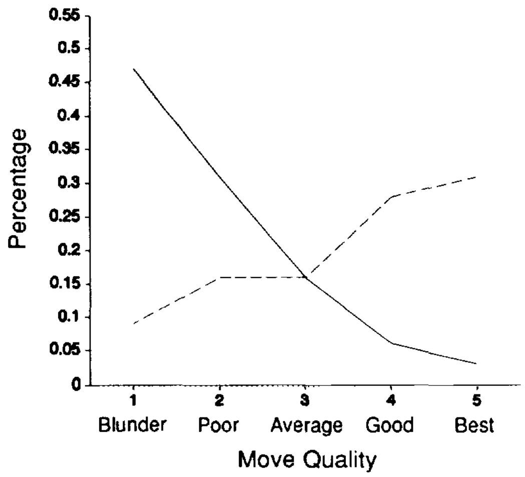

FIG. 1. (—) All legal moves; (– – –) first move generated.

TABLE 1

**Frequency of Subjects' First Generated Move vs All Legal Moves by Acceptability Level**

| | First move generated by the subjects | All legal moves |
|---|---|---|
| Acceptable moves (Grand Master points were awarded) | 41 moves | 20 moves |
| Unacceptable moves (No Grand Master points were awarded) | 23 moves | 104 moves |

*All moves generated.* Inspection of all the moves generated by subjects reveals a trend similar to that seen with the first move. Using the subjective ratings, the percentage of inferior moves (i.e., ratings of 1 or 2) was 75.8% for the legal moves, compared to 38.7% for the actual moves generated. Using the Grand Master ratings of move quality, Table 2 shows the trend for all the moves generated by the subjects, compared to all the legal moves. A $\chi^2$ analysis of the observed and expected frequencies was significant, $\chi^2$ = 189.60, $p <$ .0001. These data support the prediction generated by the RPD model that experienced chess players would generate acceptable rather than random moves to consider. The findings are also consistent with the protocol data collected by de Groot (1946/1965).

In watching the subjects, it became clear that even if the first move considered was weak, subjects would quickly dismiss it and generate another, so that there was not a great penalty for generating a poor move as the first one. This point is significant because persons acting in this way still appeared to be using a more efficient approach than generating a large or moderately-sized set of potential moves and evaluating these by comparing their strengths and weaknesses.

It is possible that the data collection procedure masked the actual move generation process, since subjects may have hesitated about reporting moves that they quickly realized were inadequate. We do not be-

TABLE 2

**Frequency of Subjects' Generated Moves vs All Legal Moves by Acceptability Level**

| | Moves generated by subjects | All legal moves |
|---|---|---|
| Acceptable moves (Grand Master points were awarded) | 139 moves | 20 moves |
| Unacceptable moves (No Grand Master points were awarded) | 145 moves | 104 moves |

lieve this happened since the subjects were carefully instructed to provide stream-of-consciousness reports, and since they did not appear to be inhibiting poor moves. Furthermore, subjects occasionally generated illegal moves, providing additional evidence against masking of the option generation process. (The illegal moves were usually of the type whereby a piece that was protecting a King was moved before the subject realized this was not allowed.) Finally, there were occasions where subjects generated inferior moves and immediately reacted to the blunder, showing no trace of inhibiting the move report.

### *H2: Size of the Option Set*

Our second hypothesis was that the players would generate a relatively small number of options. The data appear to confirm this hypothesis: the subjects generated an average of 4.63 moves for each board shown ($n = 64$ cases). The mean number of moves available for each board was 31. This suggests that the subjects were able to find a satisfactory move without widely sampling the range of available options. These data are also consistent with those reported by de Groot (1946/1965).

We also examined the data for differences across skill level. The Class A players generated fewer moves than the Class C players prior to choosing one they would play, $x = 4.5$ versus $x = 4.75$, and they chose a move generated earlier in the option generation sequence, $x = 2.06$ versus $x = 2.25$, but these differences were not significant.

### *H3: The Effect of Skill Level*

*Move selected.* We found only a small impact of skill level on our dependent measures. The Class A players tended to show higher scores, but these were not greatly different from the scores for the Class C players. The self-ratings are of little interest here, since weaker players would be likely to see mediocre moves as strong. Therefore, we looked only at the objective Grand Master ratings. The Class A players *selected* stronger moves than the Class C players, $x = 6.25$ ($n = 32$ cases) versus $x = 5.125$ ($n = 32$ cases), but this was only marginally significant, $F = 3.182$, $p = .09$.

*First move generated.* The average Grand Master ratings on the first move were also higher for the Class A players, $x = 3.75$ ($n = 32$) versus $x = 3.53$ ($n = 32$), but the difference was not significant. It is possible that the positions used did not challenge the players enough to expose differences between the two groups. However, we believe that these were challenging positions since only 6/32 times or in 19% of the cases did a Class A player generate a move rated 10 as the first move, and for the Class C players, there were only two initial moves with a rating of 10 out of 32 possibilities (6% of the cases).

*All moves generated.* For *all* moves generated, there was a trend toward higher rated moves for the Class A players. The average move generated by the Class A players was awarded $x = 3.46$ Grand Master points ($n = 32$) versus $x = 2.75$ points for the Class C players ($n = 32$), but again, these differences were not significant. It should be noted here, that while this difference is not statistically significant, it may be enough to win a game. It is possible that small, cumulative differences can add up, so that a rated difference of just .25 per move can become a 2.5 point difference over the course of 10 moves, and perhaps a 10 point difference over the course of a game, the impact of one outstanding move versus an unacceptable move. This might be enough to decide many games, but it is not large enough to be detected in a study such as this.

Despite the difficulty in demonstrating a difference between the two groups, we believe that a difference does exist. For example, the highly skilled group was clearly better able to recognize superior moves during the move rating phase of the experiment. The Class C players rated all moves higher regardless of quality (1.83 vs 1.96, $n = 992$ observations). However, when the Class A players did rate a move higher than the Class C players, it was often a move that had been awarded Grand Master Points (12/23 or 52%). For those moves where the Class C players had a higher average rating, GM points were awarded much less frequently (8/66 or 12%). A differences between proportions test was significant, $z = 3.93$, $p < .01$ (Freund—Modern Elementary Statistics, p. 317).

### *H4: The Effect of Context*

*First move generated.* In order to understand context effects, some features of the boards must be described. Board A (the only configuration from a middle game position) was fairly deceptive in that there was a clear focus on the board, and a clear move to be examined, but only after some study did the move turn out to be inadequate. Therefore, for the first move generated, none of the Class A players received any Grand Master points on this board, and only one of the eight Class C players received any points.

Context was defined as the prior familiarization with a board by beginning several moves in advance of the test position. For the self-ratings, there was a trend showing the importance of context, with the average first move generated in the context condition rated 3.81, compared to 3.31 for the no-context condition, but this difference was not significant, $F = 2.420$, $p = .13$. As would be expected, Board A, with its deceptive dynamics, accounted for most of this trend. For the other three boards, there was no familiarization effect. When the Grand Master ratings are contrasted for the context and no-context conditions, there was a trend towards higher ratings with prior familiarization, $x$ = 4.03 ($n$ = 32) versus $x$ = 3.25 ($n$ = 32), but this was not significant.

## DISCUSSION

The results provide clear support for the hypothesis that decision makers can generate reasonable options initially, making use of their experience to guide the option generation process. The data indicate that hypothesis generation takes place in an ordered, not random, fashion. There was no evidence that the subjects were randomly selecting options from the pool of legal moves. Moreover, both the subjective and objective move ratings showed that the initial moves generated were acceptable. In addition, the results indicate that the subjects did not need to generate a large number of options before finding one they would play.

These results are consistent with the work of Gettys (Gettys & Fisher, 1979) who found that subjects seriously consider only those hypotheses that are "leading contenders." Gettys also found that subjects from both student and expert populations tended to grossly overestimate the completeness of hypothesis sets and often failed to generate important hypotheses (Gettys, *et al.*, 1987). Similarly, in the study reported here, subjects generated the optimal move only 50% of the time ($n$ = 64 cases). The results of the study also suggest how Gettys' subjects might have been able to make use of such a strategy. In domains where subjects are able to generate reasonable options initially, it is appropriate to proceed with those options, particularly in domains which include time pressure. This perspective of effective hypothesis generation is in direct conflict with the advice of Janis and Mann (1977), who suggested that improving decision-making performance should be accomplished by "thoroughly canvassing a wide range of alternative courses of action" (p. 371). It is our view, based on our findings, that this advice may not be generally applicable.

The final two hypotheses regarding the impact of skill level and context did not reach significance, yet may still warrant further study. While none of the tests reported here reached the .05 criterion, each nonsignificant trend occurred in the predicted direction. One hypothesis that may explain the lack of a skill level effect is that the quality of the options generated may not be a distinguishing characteristic of skill for players above a certain strength. Evaluative skills may be a more important component in separating players at higher skill levels. This hypothesis is consistent with Holding's (1985) work on the SEEK model. A second possibility would be that the boards were not sufficiently challenging to differentiate the skills of the players, i.e., a ceiling effect. However, this did not appear to be the case. As we discussed earlier, subjects at both skill levels failed to consistently generate or select the best move which suggests that no ceiling effect was occurring. Given the observed trend, we believe that the most likely hypothesis is that stronger players do, in fact, generate better moves to consider. A similar trend was found by de Groot; in his board position A, 8/9 Grandmasters and Masters identified the best move to consider, whereas only 2/7 Experts, Class A and Class C players generated the move for consideration.

Regarding our third hypothesis that context will improve the quality of options generated; it appears that there may be an interaction between the clarity of the board position and the size of the context effect observed. For example, Board A, taken from a middle game, showed a strong difference in the context versus no-context condition for players at both skill levels. This position was the most ambiguous in terms of immediately recognizable goals and patterns. The other positions, with their more straightforward dynamics, did not elicit strong context effects from the players. Therefore, future studies should be able to show strong context effects by using positions with ambiguous goals or could diminish context effects by using positions with less complex dynamics.

An overriding conclusion of the study is that even moderately experienced people can generate feasible options, and would be best served with a serial generation/evaluation strategy, whereas inexperienced people might be better served by considering large sets of options. However, it is not clear that novices are more effective at performing analytical evaluations of large option sets than at generating adequate options. Therefore, more work needs to be done in terms of recommending option evaluation strategies for novices.

In this paper, we have distinguished between a filtering model, in which a large or moderately-sized option set is trimmed down to just a few options that can be analytically evaluated, and a generative model, in which a decision maker can generate a plausible option as the first one considered. Our data show that ordered option generation can occur. What processes could result in ordered option generation, in which the first move generated is acceptable? We suggest that a top-down and bottom-up search may be involved, in which goals become successively refined and increasingly specific as the decision maker assesses the situation, while at the same time situational features are identified and evaluated for the potential to lead to worthwhile outcomes. For example, noting that White has a Rook that is pinned directs exploration of using this weakness. Thus, the goal of capturing White's pieces is transformed into the goal of capturing the Rook, and then finally exchanging a Bishop (which has a lower value) for the Rook. When the goal coincides with the affordances of the position, a move is generated for consideration. This move still needs to be evaluated, since it may have other positive and negative consequences. Whatever the mechanism, the finding of ordered option generation illustrates the importance of an experience base for efficient decision making, particularly under time pressure. This skill in option generation suggests that it may be useful to reevaluate the costs and benefits of normative decision strategies.

## REFERENCES

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

de Groot, A. D. (1946/1965). *Thought and choice in chess* (1st ed.). New York: Mouton.

de Groot, A. D. (1987). Some benefits of advances in computer chess. *ICCA-Journal*, 10(2), 72–77.

Gettys, C. F., & Fisher, S. D. (1979). *Hypothesis plausibility and hypothesis generation*. New York: Academic Press.

Gettys, C. F., Pliske, R. M., Manning, C., & Casey, J. T. (1987). An evaluation of human act generation performance. *Organizational Behavior and Human Decision Processes*, 39, 23–51.

Holding, D. H. (1985). *The psychology of chess skill*. Hillsdale, NJ: Erlbaum.

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict, choice, and commitment*. New York: The Free Press.

Keller, L. R., & Ho, J. L. (1988, September). Decision problem structuring: Generating options. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(5).

Klein, G. A. (1989). Recognition-primed decisions. In W. B. Rouse (Ed.), *Advances in man–machine system research*, (Vol. 5, pp. 47–92). Greenwich, CT: JAI Press, Inc.

Serfaty, D., Entin, E. B., & Riedel, S. L. (1991). *The role of uncertainty in the generation, exploration, and implementation of tactical options*. Prepared for BRG Symposium on Command & Control Research. Washington, DC, June 24–25.

Received: April 15, 1993