# EVALUATING EXPERTISE IN A COMPLEX DOMAIN—MEASURES BASED ON THEORY

Jean MacMillan
Eileen B. Entin
Daniel Serfaty
ALPHATECH, Inc.
50 Mall Rd.
Burlington, MA 01803

## ABSTRACT

Human factors practitioners are often concerned with defining and evaluating expertise in complex domains where there may be no agreed-upon expertise levels, no single right answers to problems, and where the observation and measurement of real-world expert performance is difficult. This paper reports the results of an experiment in which expertise was assessed in an extremely complex and demanding domain—military command decision making in tactical warfare. The hypotheses of the experiment were: 1) command decision-making expertise can be recognized in practice by domain experts; 2) differences in the command decision-making expertise of individuals can be identified even under conditions that do not fully replicate the real world; and 3) observers who are not domain experts can recognize the expert behaviors predicted by a mental-model theory about the nature of expertise. In the experiment, the expertise of military officers in developing tactical plans was assessed independently by three "super-expert" judges, and these expertise-level ratings were correlated with independent theory-based measures used by observers who were not domain experts. The results suggest that experts in a domain have a shared underlying concept of expertise in that domain even if they cannot articulate that concept, that this expertise can be elicited and measured in situations that do not completely mimic the real world, and that expertise measures based on a mental-model theory can be used effectively by observers who are not experts in the domain.

## INTRODUCTION

Human factors practitioners deal with the selection, training, and support of experts in a variety of human endeavors. An understanding of expertise is needed to select individuals with the potential to become experts, to train non-experts, and to design and evaluate decision-support systems for both experts and non-experts. Each of these application areas can benefit from a deeper understanding of the nature of expertise and from improved methods for evaluating the expertise of individuals.

Much previous research on expertise (Chi, Glaser & Farr, 1988; Shanteau, 1992) has examined domains where there are well-defined levels of expertise (e.g., chess) or where there is a relatively linear progression through certain types of training and experience that produce expertise (e.g., medicine). Expertise is often studied using problems that have "right" answers, e.g., physics problems that have correct solutions. The application areas of interest for human factors practitioners may involve expertise that is not so well defined, however. For many complex domains there are no agreed-upon definitions for expertise levels, there are no single right answers to real problems, and the observation and measurement of expert performance in real-world situations may be difficult or even dangerous. The human factors practitioner may be asked, for example, to design a decision support system that "improves performance," where there is little agreement as to who is an expert and no criteria for assessing the quality of responses.

This paper reports the results of an experiment in which expertise was assessed in an extremely complex and demanding domain—military command decision making in tactical warfare. This is an area in which there are no accepted criteria for identifying the expert and no simple right answers for most situations, although some solutions may be greatly superior to others. It is also an area in which the observation of expert performance in real-world battlefield conditions is impossible. The hypotheses of the experiment were: 1) command decision-making expertise can be recognized in practice by domain experts, even though there is no accepted definition of this expertise; 2) differences in the command decision-making expertise of individuals can be identified even under simplified laboratory conditions that do not fully replicate the real world; and 3) observers who are not themselves domain experts can recognize expertise based on certain behaviors of the expert, as predicted by a theory about the nature of that expertise.

## METHOD

Twenty-six military officers served as subjects for the study. Based on their rank, all of the officers were expected to have some tactical expertise, but the level of that expertise was expected to vary based on ability and/or experience. Subjects were presented with written materials and maps that described three tactical situations set in the Persian Gulf. Acting as division or brigade commanders, subjects were asked to take appropriate action in the situations. This included developing a tactical plan, preparing a written statement of intent, and preparing written messages to be sent to the subordinate and lateral commanders responsible for carrying out that intent. Subjects were allowed to ask questions about the situation, and were asked to explain the rationale for their courses of action (COAs). All experiment sessions were videotaped.

The tactical decision-making expertise shown by each subject in each situation was judged independently by three "super-experts" in the domain. These judges were retired three- and four-star generals with extensive tactical experience. The judges were not supplied with a definition of expertise; instead, we asked them to define it for themselves and then assess it for each of the subjects. The judges rated the written materials produced by the subjects (intent statements and messages) and, independently, rated the videotapes in which the subjects asked questions and

explained their plans. Finally, each judge was asked to rate each subject's overall level of expertise as manifested in all of the situations—a single expertise level for each subject. We then computed the correlations between the judges' assessments of a subject's expertise to determine the extent to which the three judges agreed in their concept of expert performance as applied to specific individuals.

Two observers who were not domain experts also rated videotaped behavior using a set of measures based on a theory of command decision-making expertise. The theory predicted that more-expert subjects would manifest certain types of behavior (explained below) as they developed their plans. Some of the measures coded by the observers were based on the presence or absence of a behavior, while others required the observers to rate degree or amount, e.g., the amount of detail in a plan. Note that the ratings made by the observers did not require them to judge the "quality" or correctness of the subjects' responses and plans, and that the observers worked independently of the judges and had no access to the judges' expertise-level ratings. Subjects in the experiment also completed questionnaires concerning their attitudes and perceptions; these questions were suggested by the command-expertise theory.

After all observations and ratings had been completed, we computed the correlation of the judges' expertise ratings with the observers' ratings and the subjects' responses in order to determine which of the measures suggested by the theory were in fact associated with expertise. If the theory accurately describes the nature of command decision-making expertise, then measures based on that theory should be correlated with the expertise level of each subject as it was assessed by the judges.

## RESULTS

### Reliability of Judges' Ratings

A primary premise of the investigation was that command decision-making expertise can be identified and reliably assessed, even though there is no accepted definition of this expertise. If our hypothesis is correct, then the three judges' overall expertise assessments should be highly correlated. Table 1 shows the correlation matrix of the three judges' average expertise ratings for each subject (computed from each judge's ratings of the written products and the videotapes for the individual situations), and the coefficient alpha (Cronbach, 1970; Nunnally, 1967) derived from the correlation matrix. All of the correlations are significantly different from zero ($p<.05$). The coefficient alpha, a measure of internal consistency and an important form of reliability, is .81. This is high, implying high reliability or consistency among the judges. Moreover, it has been shown (Nunnally, 1967) that the square root of coefficient alpha, which in this case is .90, represents the correlation with the true scores (expertise) could they be known.

Table 1. Correlation of Expertise Ratings for Three Judges

| | Judge 1 | Judge 2 |
|---|---|---|
| Judge 1 | — | |
| Judge 2 | .58 | — |
| Judge 3 | .63 | .73 |

Coefficient Alpha = .81

### Measures Based on a Theory of Expertise

The measures used in the experiment were based on a theory that draws together the cognitive science literature on expertise with the results of a series of semi-structured interviews with expert commanders (Serfaty, MacMillan, and Deckert, 1991). The theory is based on the cognitive-science concept of "mental models" as internal representations of the external world (Gentner and Stevens, 1983; Johnson-Laird, 1983). We suggest that an expert commander has a mental model of the tactical situation that differs in measurable ways from that of a less-expert commander.

Figure 1 summarizes the major components of tactical decision-making expertise suggested by the theoretical framework, lists the measures associated with each component of this theory, and shows which of the measures used in the experiment were found to be significantly correlated with expertise level as assessed by the judges (correlations with a p value of less than 0.10 are highlighted) as well as the direction (positive or negative) of the correlation. Roughly half of the measures suggested by the theory were found to be significantly correlated with expertise level.

According to the theory, the expert maintains an extensive store of specific experiences in memory, and is able to rapidly retrieve an initial schema and a possible plan of action when confronted with a new situation. This component of the theory was only partially supported by the results. The more-expert subjects generated initial COAs that were more detailed than those of the less-expert subjects, but there was no indication that the experts required less time to generate these initial COAs. Also, more-expert subjects were no more likely than less-expert subjects to report that the new tactical situation was similar to their past experiences.

The hypothesis that the expert is able to focus immediately on critical unknowns and to ask the right questions was supported by significant correlations for two measures: the number and criticality of the questions asked (combined into one weighted measure), and the use of the answers to questions in developing the COA. The criticality of questions was based on ratings of a list of possible questions by the domain-expert judges.

The hypothesis that the expert builds and uses a richer (more complex) mental model was supported by all five of the measures used, each of which is an indirect indication of the creation and use of mental models. Experts were more likely to develop COAs that took account of the sequencing and timing of events, indicating that they had a better mental representation of these complex relationships than non-experts. They were also more likely to use the map as a visualization tool while planning, indicating that they were considering the effects of terrain and distance on possible movements, another element of complexity. Based on the questionnaire responses, the more-expert subjects perceived the initial tactical situation as *more complex* than did the less-expert subjects, and felt that the time and information provided was less adequate. We conclude that, based on the same information, the expert "sees" more complexity in the situation than the non-expert. This is consistent with the idea of a richer mental model.

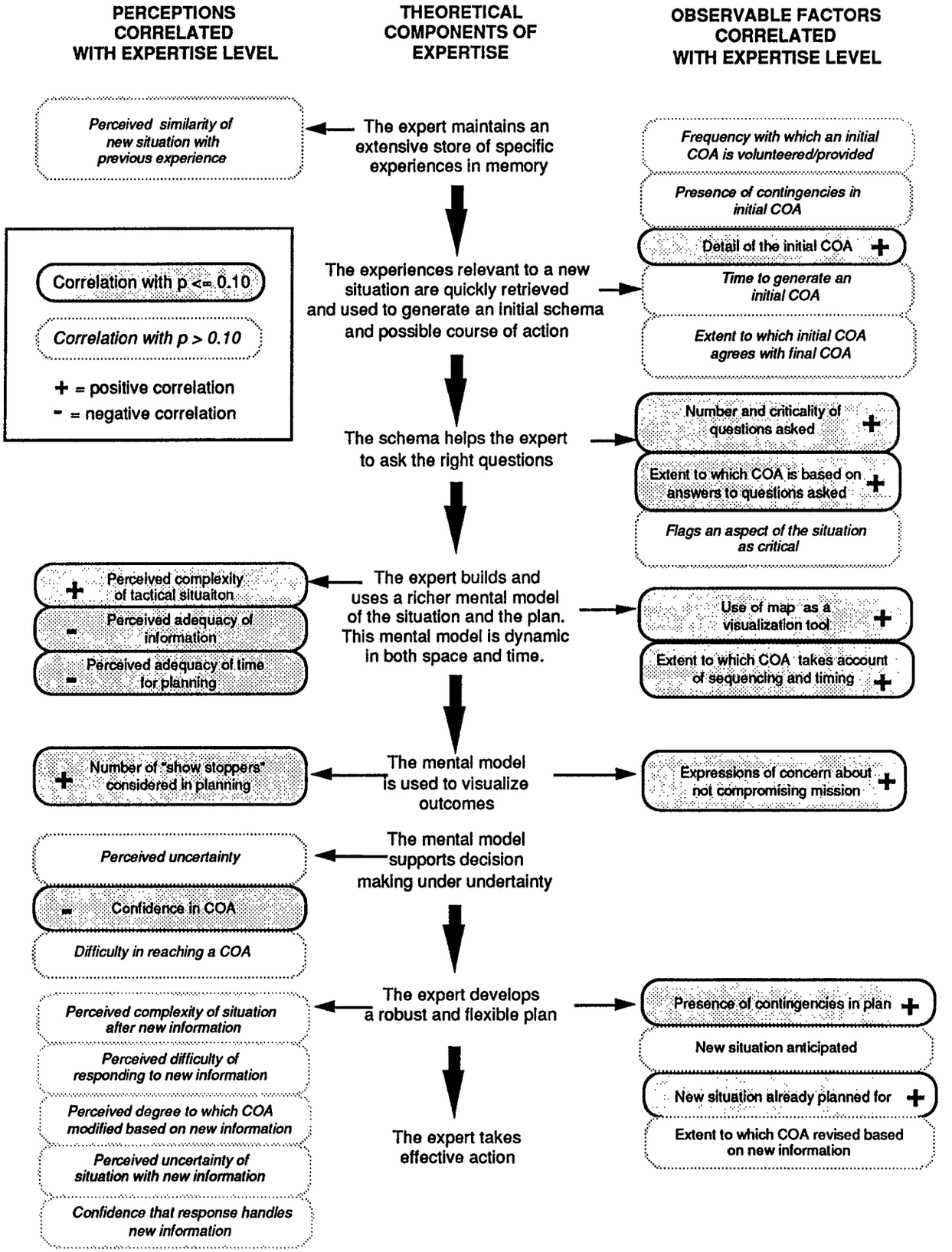


Figure 1. Correlation of Measures Based on Theory with Expertise Level (Direction of Correlation is Shown for Correlations with p<0.10).

The theory suggests that the expert uses the mental model to visualize outcomes, and that the mental model allows the expert to act under uncertainty. A number of measures support this part of the theory. Experts were more likely to mention their concern about possible compromises to the overall mission outcome, and, when asked, volunteered more "show stoppers" that they had considered when formulating their plans. Interestingly, the more-expert subjects were *less confident* about the outcomes of their COAs than the less-expert subjects. Experts seem better able to "see" what could go wrong with their plans, which is also consistent with the idea of the creation and use of a richer mental model to visualize outcomes.

The expert is able to produce a more robust and flexible plan of action because he or she has already visualized what could go wrong, and has planned for it. This was supported in the experiment by two measures: the presence of contingencies in the plans developed by the more-expert subjects, and the extent to which changes in the tactical situation had already been planned for in the expert's COA.

## DISCUSSION

We conclude that there is an underlying quality called military command decision-making expertise, that this quality is associated with individuals, that this expertise can be elicited using written materials and maps to pose a tactical problem in a laboratory setting, and that this expertise can be reliably measured by well-qualified judges. If any of these factors were not true, we would be hard pressed to explain the high agreement among the judges. While we cannot prove that our mental-model theory of expertise is correct, there is much correlational evidence to support it. Also, the theory was able to suggest process measures of expertise that are not dependent on the observer's domain knowledge. Observers who were not domain experts were able to detect the presence or absence of expert behaviors and to rate the degree to which expert characteristics were present.

The results of the experiment provide encouragement for researchers and practitioners working to understand and measure expert performance in complex domains. The results suggest that experts in a domain may have a shared underlying concept of expertise in that domain even if they cannot articulate that concept, that this expertise can be elicited and measured in situations that do not completely mimic the real world, and that certain expert behaviors in complex domains can be recognized by observers who need not be domain experts.

## ACKNOWLEDGMENTS

This research was sponsored by the Army Research Institute Field Unit, Fort Leavenworth, Kansas and benefited greatly from the assistance and support of Mr. Rex Michel of ARI. The methodology and results of the research are reported more fully in Deckert, Entin, Entin, MacMillan, and Serfaty (1993).

## REFERENCES

Chi, M.T.H., Glaser, R., and Farr, M.J. (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L.J. (1970). *Essentials of psychological testing*. Third Edition. New York: Harper and Row.

Deckert, J.C., Entin, E.B., Entin, E.E., MacMillan, J., and Serfaty, D. (1993). Military command decision-making expertise—Annual interim report. TR-568-1, Burlington, MA: ALPHATECH, Inc.

Gentner, D. and Stevens, A.L. (Eds.) (1983). *Mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P.N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Nunnally, J.C. (1967). *Psychometric theory*. New York: McGraw Hill.

Shanteau, J. (1992). The psychology of experts. In G. Wright and F. Bolger (Eds.), *Expertise and decision support*. New York: Plenum Press.

Serfaty, D., MacMillan, J., and Deckert, J.C. (1991). Toward a theory of tactical decisionmaking expertise. TR-496-1, Burlington, MA: ALPHATECH, Inc.