Measuring Team Knowledge During Skill Acquisition of a Complex Task

Nancy J. Cooke
New Mexico State University

Preston A. Kiekel
New Mexico State University

And

Erin E. Helm
Austin Usability

Running Head: Measuring Team Knowledge

Contact Information:

Nancy J. Cooke
Department of Psychology
New Mexico State University
Las Cruces, NM 88003

Voice: 505-646-1630
Fax: 505-646-6212
Email: cooke@crl.nmsu.edu

ABSTRACT

The design of team training programs and other team interventions could benefit from an understanding of team cognition. The research presented in this article evaluates methods for eliciting and assessing team knowledge during acquisition of a complex task. Knowledge measures are evaluated in terms of their ability to predict team performance and also in terms of how they reflect skill acquisition. The study was conducted in the context of a synthetic three-person team task that involved operating an uninhabited air vehicle. Eleven teams of three Air Force ROTC cadets participated in three experimental sessions lasting from three to six hours. During these sessions teams were trained on the task and were observed as they performed ten 40-minute missions. During the missions, team performance and team process behaviors were measured, as well as the fleeting team knowledge associated with situation awareness. In addition, long-term team knowledge regarding both taskwork and teamwork were measured off-line in four sessions. Results indicated that teams reached asymptotic performance on this task after 1.5 hours of individual training and four 40-minute team missions. This skill acquisition was paralleled by improvements in team situation models, teamwork knowledge and to a lesser extent, team process behaviors. Taskwork relatedness ratings measured at both the individual and team level were good predictors of team performance and indicated that high performing teams had more knowledge of the task from the perspective of other team members, as opposed to lower performing teams. These measures help reveal the knowledge underlying team behavior, and thus have implications for team training and other interventions.

MEASURING TEAM KNOWLEDGE DURING SKILL ACQUISITION OF A COMPLEX TASK

Technological developments in the military and elsewhere have transformed highly repetitive manual tasks, requiring practiced motor skills into tasks that require cognitive skills often related to overseeing new technology such as monitoring, planning, decision making, and design (Howell & Cooke, 1989). As a result, a full understanding of many tasks, at a level required to intervene via training or system design, requires an examination of their cognitive underpinnings. Additionally, the growing complexity of tasks frequently surpasses the cognitive capabilities of individuals and thus, necessitates a team approach. For instance, teams play an increasingly critical role in complex military operations in which technological and informational demands necessitate a multioperator environment (Salas, Cannon-Bowers, Church-Payne, & Smith-Jentsch, 1998).

Whereas the team approach is often seen as a solution to cognitively complex tasks, it also introduces an additional layer of cognitive requirements that are associated with the demands of working together effectively with others. Team members need to coordinate their activities with others who are working toward the same goal. Team tasks often call for the team to detect and recognize pertinent cues, make decisions, solve problems, remember relevant information, plan, acquire knowledge, and design solutions or products as an integrated unit. Therefore, an understanding of team cognition, or what some have called the new "social cognition" (Klimoski & Mohammed, 1994), is critical to understanding much team performance and intervening to prevent errors or improve productivity and effectiveness.

The assessment and understanding of team cognition (i.e., team mental models, team situation awareness, team decision making) requires psychometrically sound measures of the constructs that comprise team cognition. However, measures and methods targeting team cognition are sparse and fail to address some of the more interesting aspects of team cognition (Cooke, Salas, Cannon-Bowers, & Stout, 2000). As measures of team cognition are developed, they can be used to better understand team cognition.

The research presented in this article evaluates methods for eliciting and assessing team knowledge during acquisition of a complex task. Knowledge measures are evaluated in terms of their ability to predict team performance and also in terms of how they reflect skill acquisition. The knowledge measures that are evaluated here attempt to address some of the shortcomings of the current methodologies. In the following section team knowledge is defined in terms of a framework that also identifies some of these shortcomings.

Team Cognition and Team Knowledge

Salas, Dickinson, Converse, and Tannenbaum (1992) define *team* as "a distinguishable set of two or more people who interact dynamically, interdependently, and adaptively toward a common and valued goal/object/mission, who have each been assigned specific roles or functions to perform, and who have a limited life span of membership" (p. 4). Thus, teams, unlike some groups, have differentiated responsibilities and roles (Cannon-Bowers, Salas, & Converse, 1993). This division of labor is quite common and enables teams to tackle tasks too complex for any individual.

Interestingly, this feature is also one that has been neglected by current measurement practices (e.g. Langan-Fox, Code, & Langfield-Smith, 2000).

There has been significant theoretical work delineating cognitive constructs such as team decision making, shared mental models, and team situation awareness (Cannon-Bowers, et al., 1993; Orasanu, 1990; Stout, Cannon-Bowers, & Salas, 1996). It is assumed that with an understanding of these constructs, training and design interventions can target the cognitive underpinnings of team performance. Also, the hypothesized relation between team cognition and team performance suggests that team performance can be predicted from an assessment of team cognition. The ability to predict team performance from team cognition suggests that team performance may ultimately be assessed indirectly through cognitive measures independent of the task, thereby circumventing the need for teams to perform in less than optimal settings (e.g., hazardous or high-risk environments).

The research presented in this article focuses on team knowledge. Parallel to research on individual expertise (e.g., Chase & Simon, 1973; Glaser & Chi, 1988), accounts of effective team performance highlight the importance of knowledge, or in this case, team knowledge. For instance, Cannon-Bowers and Salas (1997) have recently proposed a framework that integrates many aspects of team cognition in the form of teamwork competencies. They categorize competencies required for effective teamwork in terms of knowledge, skills, and attitudes that are either specific or generic to the task and specific or generic to the team. Similarly, a team's understanding of a complex and dynamic situation at any one point in time (i.e., team situation awareness) is supposedly

influenced by the knowledge that the team possesses (Cooke, Stout, & Salas, 1997; Stout, et al., 1996).

Figure 1 presents team knowledge in a framework that serves to better define team knowledge, and at the same time to identify some issues in its measurement. Traditional measures of team knowledge operate at the collective level, eliciting knowledge from individuals on the team and then aggregating the individual results to generate a representation of the collective knowledge of a team. Although collective knowledge should be predictive of team performance, it is devoid of the influences of team process behaviors (e.g., communication, coordination, situation awareness), analogous to individual cognitive processes. The process behaviors transform the collective knowledge into effective knowledge. This effective knowledge is what we describe as the holistic level and is associated with actions and ultimately, with team performance. Also, note in Figure 1 that team knowledge consists of background knowledge that is long-lived in nature, as well as more dynamic and fleeting understanding that an operator has of a situation at any one point in time. Measures of team knowledge have focused primarily on the former, at the expense of the latter.

[Insert Figure 1 about here]

Other measurement issues include the fact that traditional measures of team knowledge, focusing primarily on the similarity of team members' knowledge, tend to assume homogeneity with respect to team composition as opposed to the heterogeneous backgrounds suggested by the definition of team offered above. In addition, methods for

aggregating individual knowledge to derive collective knowledge are worthy of further study (e.g., the social decision scheme literature (Hinz, 1999; Davis, 1973)), as are measures that target different types of team knowledge (e.g., strategic, declarative, procedural knowledge or task vs. team knowledge). Finally, measures of team knowledge could benefit from a broader application of various knowledge elicitation techniques, procedures to better automate the measures and embed them within task contexts, and evaluations of measures in terms of validity and reliability. Details of these issues are described in Cooke, et al. (2000).

The team knowledge measures that are evaluated in this paper attempt to address some of these measurement issues. In particular, measures target both long-term background knowledge of both taskwork and teamwork varieties, as well as knowledge associated with the more dynamic situation models (i.e., a team's understanding of the situation at one point in time). In addition, one of the elicitation methods targets the holistic level and a set of team knowledge metrics is used to take into account the heterogeneous nature of the teams.

The Synthetic Task Environment

We assume that the task and surrounding environment are inextricably tied to team knowledge and its measurement. This makes the selection of the team task and setting critical. While field settings provide realistic opportunities for observation, they do not afford the experimental control and measurement flexibility needed for the development and evaluation of measures. Therefore, these studies were conducted in the context of an STE (Synthetic Task Environment), based on the real task of controlling a UAV (Uninhabited Air Vehicle).

Synthetic tasks are "research tasks constructed by systematic abstraction from a corresponding real-world task" (Martin, Lyon, & Schreiber, 1998, p. 123). An STE provides the context for a suite of synthetic tasks. Performance on a synthetic task should exercise some of the same behavioral and cognitive skills associated with the real-world task. This environment offers a research platform that bridges the gap between controlled studies using artificial laboratory tasks and uncontrolled field studies on real tasks or using high-fidelity simulators.

The STE used in these studies is an abstraction of the Air Force's Predator UAV operations (Cooke, Rivera, Shope & Caukwell, 1999; Cooke, Shope, & Rivera, 2000). It is implemented in the context of NMSU's CERTT (Cognitive Engineering Research on Team Tasks) Laboratory that contains hardware and software for simulating a variety of team tasks and adequate measurement and task manipulation capabilities for research on those tasks. CERTT's UAV-STE is a three-person task in which each team member is provided with distinct, though overlapping, training; has unique, yet interdependent roles; and is presented with different and overlapping information during the mission. The overall goal is to fly the UAV to designated target areas and to take acceptable photos at these areas.

The AVO (Air Vehicle Operator) controls airspeed, heading, and altitude, and monitors UAV systems. The PLO (Payload Operator) adjusts camera settings, takes photos, and monitors the camera equipment. The DEMPC (Data Exploitation, Mission Planning and Communication Operator) oversees the mission and determines flight paths under various constraints. To complete the mission, the team members need to share

information with one another and work in a coordinated fashion. Most communication is done via microphones and headsets, although some involves computer messaging.

The CERTT UAV-STE was abstracted from results of a cognitive task analysis (Gugerty, DeBoom, Walker, & Burns, 1999) of the Predator operational environment, with the goal of providing an experimenter-friendly test-bed for the study of team cognition. As a result, cognitive aspects of the task are emphasized and other task components (e.g.. the specific interface, stick-and-rudder control have been omitted). For instance, alterations in the interface enable individual team members to rapidly (within 1.5 hours) acquire the skills and knowledge needed to work as an integral part of the team. Measures taken include audio records, video records, digital information flow data, embedded performance measures, team process behavior measures, and a variety of individual and team knowledge measures.

Overview of Study

The study was conducted in the context of the team UAV task. Eleven teams of three Air Force ROTC cadets participated in three experimental sessions lasting from three to six hours. During these sessions teams were trained on the task and were observed as they performed ten 40-minute missions. During each mission team performance, team process behavior and team situation models were measured. In addition, long-term team knowledge regarding both taskwork and teamwork were measured apart from the missions in four sessions. This study was designed to evaluate a number of different approaches to measuring team knowledge and to examine the development of team performance, process, and knowledge as team skill was acquired over the ten missions. The validity of team knowledge measures was assessed in terms of

the ability of measures to predict team performance and process. In addition the patterns of acquisition of team knowledge, performance, and process were also examined. These patterns may shed light on sequential dependencies among components of team performance and in addition, provide useful information about the point at which asymptotic performance is reached in this synthetic task

METHOD

Participants

Eleven three-person teams of Air Force ROTC cadets voluntarily participated in three (3-5 hour) sessions of this study. Individuals were compensated for their participation by payment of $6.00 per person hour to the ROTC organization. In addition, the three team members on the team of with the highest mean performance score were each awarded a $50.00 bonus.

Equipment and Materials

The study took place in New Mexico State University's CERTT (Cognitive Engineering Research on Team Tasks) Lab, configured for the UAV team synthetic task described above. For most of the study, each participant was seated at a workstation consisting of two computer monitors (one View Sonic monitor connected to an IBM PC 300PL and one Cyberesearch Industrial Workstation), a Sony video monitor that could present video from a Quasar VCR, a keyboard, a keypad, and a mouse for input. Participants communicated with each other and the experimenters using David Clark headsets and a custom-built intercom system, designed to log speaker identity and time information. The intercom enabled participants to select one or more listeners by flipping toggle switches.

Two experimenters were seated in a separate adjoining room at an experimenter control station consisting of another IPB PC computer and View Sonic monitor, headsets for communicating with participants, and Panasonic monitors for video feed from ceiling-mounted Toshiba cameras located behind each participant. In addition, a fourth camera captured information from the entire participant room. From the experimenter workstation, the experimenters could start and stop the mission, query participants together or individually, monitor some of the mission-relevant displays, observe team behavior through camera and audio input, and enter time-stamped observations. Video data from cameras were recorded on a Quasar VCR. Audio data from the headsets were recorded on an Alesis digital recorder as well as to the VCR. In addition, custom software recorded communication events in terms of speaker, listener, and the interval in which the push-to-talk button on the microphone was depressed.

Custom software (seven applications connected over a local area net) also ran the synthetic task and collected values of various parameters that were used as input by performance scoring software. A series of tutorials were designed in Powerpoint for training the three team members. Two of the three Powerpoint modules were unique to each position. Custom software was also developed to conduct tests on information in Powerpoint tutorials, to collect individual and consensus taskwork relatedness ratings, and to collect demographic and preference data at the time of debriefing.

In addition to software, some mission-support materials (rules-at-a-glace for each position, two screen shots per station corresponding to that station's computer displays, and examples of good and bad photos for the Payload Operator) were presented on paper at the appropriate workstations. Other paper materials consisted of the consent forms,

debriefing form, a checklist of skills for training, forms for experimenter recording of responses to situation model queries and observations of process behaviors, a trust survey, and teamwork and taskwork questionnaires.

Measures

Performance, process, and knowledge measures are the focus of this paper, though demographic, preference, trust, video, and communication data were also collected; they are not addressed in this article.

Team performance was measured using a composite score based on the result of mission variables including time each individual spent in an alarm state, amount of fuel used, amount of film used, number of targets successfully photographed, and number of critical waypoints visited. Penalty points for each of these components were weighted a priori in accord with importance to the task and subtracted from a maximum score of 1000.

Team process behavior was scored independently by each of the two experimenters. For each mission the experimenters observed team behavior and responded yes or no to each of nine team process behaviors based on whether that behavior did or did occur at designated event-triggers in each mission. Team process was simply the proportion of the nine process questions that were observed by each experimenter. The process behaviors and triggers are presented in Table 1. Similar questions were used for Missions 7 and 10, with different event details to accommodate the different scenario.

[Insert Table 1 about here]

Team knowledge was measured using several different methods outlined in Table 2. Team situation models were measured using three SPAM-like (Durso, Hackworth, Truitt, Crutchfield, Nikolic, & Manning, 1998) queries administered during the mission. Query order and the time (in 5 minute increments) at which queries began were both randomly determined without replacement. One of the experimenters administered the queries to each team member in turn during the five-minute interval. Order in which team members were queried was also random. The three queries asked (1) a prediction regarding the number of targets out of nine successfully photographed by the end of the mission; (2) the team member or members that they would communicate with next and the topic of that communication; and (3) the number of targets out of nine successfully photographed thus far. The experimenter also recorded the correct response to these queries once known. Responses to the queries were scored for accuracy, as well as intrateam similarity. Team accuracy scores were based on the average accuracy of team members, as scored using the experimenter-generated key. For the second query, this was simply the proportion of correct responses (1 or 0) averaged across the three team members. For the first and third queries, this was the absolute value of the deviation from correct, divided by 9 possible targets and subtracted from 1. For the first and last queries, team similarity was the average of all the pairwise similarities (i.e., converse proportions of absolute deviations) of the three team members. Intrateam similarity was not meaningful for the second query.

[Insert Table 2 about here]

Longer-term team knowledge was measured in four separate sessions by four methods: teamwork questionnaire, taskwork questionnaire, taskwork ratings, and taskwork consensus ratings. The teamwork questionnaire consisted of a three-part question in which the individual was asked to indicate if directed pairs of team members (e.g., AVO → PLO) pass information in the specified direction. The second part of the question asked them to identify the nature of the information for those communication links identified. The third part asked them to consider any sequential constraints in the timing of the information.

The taskwork questionnaire asked individuals to analyze the task starting with the main goal and breaking this up into subgoals and tasks. The next part of this questionnaire asked individuals to associate team roles with each of the tasks and then to indicate any sequential constraints in tasks.

The taskwork ratings consisted of eleven task related terms: altitude, focus, zoom, effective radius, ROZ entry, target, airspeed, shutter speed, fuel, mission time, and photos. All possible pairs of these terms were presented in one direction only, one pair at a time. Pair order was randomized and order within pairs was counterbalanced across participants. Each team member rated the relatedness of each pair on a 1-5 scale with anchors that ranged from slightly related to highly related. There was also an option of unrelated.

Taskwork consensus ratings consisted of the same pairs as taskwork ratings (randomly presented), however the ratings were entered as a team. For each pair, the rating entered in the prior session by each team member was displayed on the computer

screen of that team member. The three team members discussed each pair over their headsets until consensus was reached.

The longer- term knowledge measures were each scored for accuracy and intrateam similarity. Individual accuracy scores and pairwise measures of response similarity were averaged across team members. For the two rating tasks, data were submitted to KNOT (using parameters *r*=inf. And *q*=n-1) in order to generate Pathfinder networks (Schvaneveldt, 1990). These networks reduce and represent the rating data in a meaningful way in terms of a graph structure with concept nodes standing for terms and links standing for associations between terms. A referent network generated by the experimenters served as the key, and similarity of any one network to this referent in terms of the proportion of shared links was used as a measure of accuracy. In addition, the individual task ratings were scored not only against a key representing overall knowledge, but also against role-specific keys. In this way, measures of "role" or "positional" accuracy, as well as "interpositional" accuracy (i.e., interpositional knowledge (IPK) or knowledge of roles other than their own) could be determined. Team accuracy was the mean accuracy across team members. Intrateam similarity was measured using the proportion of shared links for all intrateam pairs of two individual networks (i.e. the mean of the three pairwise similarity values among the three networks).

Procedure

The study consisted of three sessions. Sessions 1 and 2 lasted approximately 5.5 hours each and were separated by a 24-48 hour interval. Session 3 lasted 3.5 hours and followed Session 2 by a lapse of 4 to 8 weeks. During this time seven of the 11 teams

participated in a team strategic training seminar offered by the ROTC for the purpose of a separate study.

In the first session the three participants were randomly assigned to one of the three task positions: AVO, PLO, or DEMPC. Team members retained these positions within the same team for the remainder of the study. The team members were given a brief overview of the study and then were seated at their workstations for training. Team members studied the three Powerpoint training modules at their own pace and were tested with a set of multiple-choice questions at the end of each module. If responses were incorrect they were instructed to go back to the Powerpoint tutorial and correct their answers. Experimenters provided assistance and explanation if their second response was also incorrect. Once all team members completed the tutorial and test questions, a mission was started and experimenters had participants practice the task, checking off skills that were mastered (e.g., the AVO needed to change altitude and airspeed, the PLO needed to take a good photo of a target) until all skills were mastered. Again, the experimenters assisted in cases of difficulty. All teams achieved the criteria for the PowerPoint and skill-based training in less than 1.5 hours total.

After a short break the first 40-minute mission began and was completed at the end of the 40-minute interval or when team members believed that the mission goals had been completed. Knowledge measures were administered on all occasions in the following order: taskwork ratings, taskwork consensus ratings, taskwork questionnaire, and teamwork questionnaire. In general, the sessions consisted of breaks, missions and knowledge sessions as presented in Table 3. Missions 7 and 10 involved the same scenario, which differed from the other 8 missions.

[Insert Table 3 about here]

RESULTS

Overview of Analyses

One team did not complete Session 3 and due to equipment malfunctions another team had no performance data recorded for Mission 10. Therefore there are performance, process, and situation model data missing for 1 or 2 teams for Missions 8 through 10.

There was adequate agreement between the two experimenters on the team process questions. Agreement between raters was assessed by computing the proportion of agreement between raters across the nine process questions for each team, each mission, and overall. Overall proportion of agreement was .9 (range from .83 to .97). Therefore, ratings were averaged for all cases in which two raters each assigned a score.

A cluster analysis of accuracy and similarity results for the three team situation model queries was used to identify meaningful groupings of the six metrics. This resulted in four clusters: (1) accuracy to Queries 1 and 3, (2) similarity for Queries 1 and 3, (3) accuracy on the to whom answer to Query 2, and (4) accuracy on the topic answer to Query 2. These were used as indices of team situation models.

Finally, due to the use of a small sample of eleven teams, extensive across-team variation, and an objective of identifying any potentially interesting measures or effects at the expense of possible Type I errors, we considered α-levels of $\leq .10$ statistically detectable. Reported correlations of team measures were also based on eleven teams (or fewer for those missions associated with the two teams with missing data) and therefore nine degrees of freedom. Thus, correlations of .52 and higher are required for two-tailed significance at the $p =$. 10 level, though we recognize that correlations somewhat lower

nonetheless predict a substantial proportion of the variance (Cohen, 1994; Wickens, 1998).

Task Acquisition

The team performance score ranged from 353 to 952 with an overall mean of 822 and standard deviation of 74.2. As might be expected and as shown in Figure 2, the standard deviation was greatest for the first and last three missions (range from 106 to 159) and was lowest for the four middle missions (range from 24 to 51). As seen in Figure 3, across the 11 teams, performance improved in general from Mission 1 (M = 510) to Mission 10 (M = 881) (t (8) = 6.70, $p < .001$), reached asymptote at Mission 4 and then dipped at Mission 8, which was the first mission after the extended break between Sessions 2 and 3. Figure 3 also shows that this drop in performance was greatest for the lowest performing teams. Interestingly, team performance did not suffer as a result of the change in scenario that occurred for Missions 7 and 10 (M = 897 for Missions 7 and 10 and M = 894 for Missions 4, 5, 6, and 9).

[Insert Figures 2 and 3 about here]

The means for the score for team process behavior revealed a pattern of acquisition similar to that for performance, but this was not statistically detectable (F (9,87) = 1.62, p =.122, η^2=.143). The biggest improvement in consecutive missions occurred between Missions 1 and 2 (.74 to .82), but it was also not detectable (t (10) = 1.38, SE = .064, p =.199). However, due to the decrease in error variance over time, a drop of the same magnitude between Missions 8 (M = .87) and Mission 9 (M = .78) was detectable (t (9) =

4.0, *SE* =.022, *p* = .003). Mean team performance and process scores across the ten missions are shown in Table 4.

[Insert Table 4 about here]

Both team situation models and teamwork knowledge as measured by the teamwork questionnaire improved with experience. For situation models, responses to the second query (concerning to whom the individual would talk to next and about what) did not change in any discernable way over time. However, the other situation model queries did change and in a way that paralleled performance (See Table 4). Accuracy on these queries generally improved from Mission 1 to 10 (.79 to .94 respectively; t (8) = 3.875, p =.005), peaked at Mission 4 (M = .94) and dropped at mission 8 (M =.89). Also there was no difference between the standard mission scenario and the new one associated with Missions 7 and 10. Intrateam response similarity for situation model Queries 1 and 3 also increased overall (Mission 1 M= .85, Mission 10 M= .94, t (9) = 4.66, p =.001), peaked at Mission 4 (M=.93), and showed no effect of novel scenarios associated with Missions 7 and 10. There was, however, no discernable drop in team situation model similarity at Mission 8.

Finally, the teamwork questionnaire showed general improvement in team accuracy across the four sessions (M = .53, .66, .71, and .65, respectively; F (3, 29) = 3.083, p = .043, η^2 = .242). Knowledge as measured by this questionnaire seemed to change most drastically between Session 1 and Session 2 (M = .66; t (10) = 2.08, p = .065) that also corresponded to the Mission 4 asymptote seen in the performance data.

How Well Do Measures Predict Team Performance?

The team process behavior measure did not correlate reliably with performance (r (11) = .132), although several of the individual questions were correlated with performance for the asymptotic missions 4 through 7.

Critical for the assessment of the validity of knowledge measures is the degree to which they correlate with measures of team. Team situation models (averaged across the 10 missions), as measured by Queries 1 and 3 (accuracy and similarity), correlated reliably with mean team performance (also averaged across the 10 missions) (r (11) = .88, $p < .0001$ and .72, $p = .013$, respectively). Multiple regression analysis indicated that most of the predictive power was derived from the Query 1 and 3 accuracy measure (t (10) = 2.91, $p = .02$).

For correlations between long-term knowledge and performance, data from Knowledge Session 1 were used because (1) with the exception of the teamwork questionnaire there was little difference across sessions, and (2) for some measures, across-team variance increased dramatically after knowledge session 1, which may indicate that participants took the knowledge task less seriously after the first session. This is especially true for the taskwork consensus ratings for which the standard deviation of the team accuracy score increased from .04 for session 1 to .13, .12, and .14 for sessions 2 to 4 respectively. Also, given that degree of across-team performance variance changed dramatically across missions (see Figure 2), correlations of knowledge with performance at each mission were computed.

In general, for the various taskwork rating metrics (except role accuracy) and to a lesser extent for taskwork consensus rating accuracy, the measures taken in Knowledge

Session 1 were significantly predictive of team performance in the first and last missions (See Table 5). At Knowledge Session 1 greater taskwork rating accuracy, IPK, and intrateam similarity corresponded to higher team performance scores for the early and late missions. Team accuracy and intrateam similarity for both teamwork and taskwork questionnaires generally failed to predict performance.

[Insert Table 5 about here]

Taskwork consensus ratings

Taskwork consensus ratings was a new method in which team knowledge was elicited at the team-level. It was assumed that this more holistic approach to measurement would capture not only the collective knowledge of the team members, but also process behaviors of the team that are used in coming to consensus on the ratings (Cooke et al., 2000). Therefore, it was hypothesized that the consensus ratings would be better predictors of team performance than the aggregate taskwork ratings. As indicated in Table 5, accuracy of this measure correlated significantly with team performance at Missions 4 and 7. Thus, the taskwork consensus ratings, although predictive of performance, did not surpass the aggregate taskwork ratings in their predictive power. However, the accuracy of the taskwork consensus ratings did correlate with the accuracy measure based on individual ratings (r (9) = .522, p = .099), as well as IPK accuracy (r(9) = .659, p = .028).

In order to identify strategies that the teams used to come to consensus in this rating task, the three individual and one team rating for each of the 55 concept pairs was

examined for each of the eleven teams. For each pair, the set of four ratings was classified according to one of five rules that mapped individual ratings onto the team rating:

1) all agreed (e.g., AVO=5, PLO = 5, DEMPC = 5, Team = 5)
2) majority (2 out of 3) rules (e.g., AVO = 4, PLO=4, DEMPC = 3, Team =4)
3) leader emerges (e.g., AVO=3, PLO=0, DEMPC=1, Team =3 or AVO=4, PLO=4, DEMPC = 2, Team =2)
4) mid rating (e.g., AVO=0, PLO=3, DEMPC=5, Team =2 or AVO=0, PLO=3, DEMPC=5, Team=3)
5) different from each, and not middle rating (e.g., AVO=5, PLO=2, DEMPC=4, Team=0)

Results of this classification are presented in Table 6. This table illustrates that most teams used strategies 2, 3, and 5 more that 1 and 4. Further, there seems to be little correspondence between the strategies that were used and team performance. Experimenters observed that there was very little communication going on during the consensus rating process. Therefore it seems that most teams assumed the strategy to go with majority rule (strategy 2) or with the single individual who claimed to have knowledge in the area (strategy 3). It is not clear why there are so many instances of strategy 5, in which the team rating is more extreme than any individual rating. However, the preponderance of this strategy could indicate that teams did not take the consensus rating task seriously.

[Insert Table 6 about here]

DISCUSSION

In general, the results of this study indicate that teams are able to reach asymptotic levels of team performance on the synthetic UAV team task after 1.5 hours of individual training and four 40-minute missions of teamwork. The fact that asymptotic performance was reached at Mission 4 could be a result of either four trials of practice or the 24-48 hour incubation period that occurred between Sessions 1 and 2, or a combination of both. Also the data indicate that the experience acquired seems to readily transfer to a novel scenario. That is, team performance did not suffer a significant decrement with the presentation of the novel scenario for Missions 7 and 10.

On the other hand, team performance did suffer from an extended break of four to ten weeks that occurred between Sessions 2 and 3, as indicated by the drop in team performance (and team situation model accuracy) at Mission 8. In fact, some of the lowest scoring teams never recovered from this drop. It is also the case that those teams with the four lowest team performance scores at Mission 8, also had relatively long breaks (8-9 weeks) between Sessions 2 and 3. Furthermore, the acquisition of team performance on the synthetic task acquisition was paralleled by changes in team situation models and tended to be preceded by process improvements, suggesting that acquisition of effective team process behavior may be a prerequisite to successful team performance and situation awareness.

Interestingly, the only noticeable knowledge changes over the four sessions occurred for responses to the teamwork questionnaire on which teams improved across the four sessions and tended to asymptote at Session 2, paralleling the fourth mission.

Thus, the team performance and team situation model asymptote appear to be paralleled by not only team process improvements, but also by an improved understanding of the teamwork aspects of the task (i.e., knowledge of the team roles and information dependencies).

In hindsight, the first knowledge elicitation session that occurred after training and Mission 1 may have been too late to detect any changes in the two taskwork knowledge measures (questionnaire and rating). Possibly, the most significant growth in knowledge of the team and its tasks occurs during the training session as the team is just learning about the mission and how they will work together. By the time the first mission is complete then, much of the team's broad knowledge is solidified. Alternatively, subtle knowledge structure refinement associated with true expert-level performance may require more experience than teams had in this study. Also, as suggested by increasing variance in some of the knowledge measures (i.e., taskwork consensus ratings), it may be the case that fatigue and boredom contributed to increased noise and lack of reliability in the other knowledge measures, masking any knowledge acquisition that was present.

Although the taskwork relatedness ratings and the taskwork consensus ratings demonstrated little improvement over time, the measures taken in the first knowledge session were predictive of team performance. Those teams with greater knowledge accuracy, IPK, intrateam similarity and consensus accuracy in the beginning tended to have higher scores on early and late missions. This pattern indicates that teams with members who understand the task from the perspective of other positions, and therefore have knowledge similar to one another, are the teams with the highest performance.

In general, the taskwork rating measures seem to be valid indicators of team knowledge, compared to the taskwork and teamwork questionnaires that failed to correlate with team performance. In addition, the more holistic taskwork consensus ratings were also predictive of team performance in some missions. The first session knowledge measures were most predictive of performance and were also associated with lower error variance compared to later knowledge sessions. This pattern again suggests that in general, the rating tasks may be most informative upon first administration.

The taskwork consensus rating task was a new measure developed in attempt to capture the holistic aspects of team knowledge that include not only the aggregate of individual team member knowledge, but also the effects of team process behaviors (see Figure 1). Examination of consensus rating strategies suggests that, quite often, if two team members rated two concepts the same, the third team member conformed to their answer. When none of the members initially agreed on a rating, another popular strategy was for one team member, usually the team member that was considered to have the most knowledge or experience with those concepts, to convince the other team members to change their ratings. Interestingly, the team generally did not just choose to rate the concepts somewhere in the middle of all of their answers (averaging), but instead went with the perceived expertise of one or more of their team members. It is difficult to explain the preponderance of strategy 5, in which teams entered a rating completely different from, and not a mid point of the three individual ratings.

Although the consensus rating task was predictive of performance and also correlated with the individually-based taskwork ratings, it did not surpass the traditional aggregate measure (i.e., taskwork ratings) in predicting performance (see Table 5). This

may call into question the value of the consensus ratings. In many scenarios, it will be more difficult to assemble the team and achieve consensus than to ask team members to rate concepts individually, to be averaged later. A counter to this argument is to note that, although the correlation between the consensus and aggregate approaches was moderately high, (r (9) = .522, p = .099), it did not approach colinearity. Hence, these two metrics tap different constructs.

There are several logical but untested explanations for the relative weakness of the taskwork consensus rating method. One possibility relates to the increasing error variance associated with rating tasks in general. The fact that the consensus ratings always followed a set of individual ratings may have exacerbated this problem for the consensus ratings. That is, teams were bored and tired and wanted to quickly finish the task. A second, and related, explanation is inability to concentrate on the consensus rating task, brought on by the pressure for off-task social interaction, coupled with the knowledge that the bonus was tied to mission performance and not the rating task. These two hypotheses are in accord with the preponderance of an apparently random social decision scheme during the consensus ratings. Perhaps teams did not take the task seriously, and simply entered ratings until the three matched.

Overall, the results of this study suggest that the team situation model queries, the teamwork questionnaire, and the taskwork rating tasks provide valid indicators of team knowledge. In particular, the team knowledge metrics used here are appropriate for teams in which members have different roles. Applying these heterogeneous metrics to the data reveal that highest performing teams have members with more knowledge of the tasks from the perspective of roles other than their own. In other words, knowing multiple

roles is better than simply knowing your own. Thus, high performing teams seemed to naturally acquire the kind of knowledge that is consistent with cross training. Measures of team knowledge provide a window to some of the cognitive the factors underlying team acquisition of a complex skill and can thus be valuable in designing and assessing knowledge-based training programs. For example, these data suggest that a team's acquisition of this task would benefit from team member cross training. The assessment of a cross training program that applied the heterogeneous knowledge metrics to the problem would provide information on the effects of the training not only on team performance, but also on team knowledge.

REFERENCES

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In J. Castellan Jr. (Ed.), Current *issues in individual and group decision making* (pp. 221-246). Hillsdale, NJ: Erlbaum.

Cannon-Bowers, J. A., and Salas, E. (1997). Teamwork competencies: The interaction of team member knowledge skills and attitudes. In O. F. O'Neil (Ed.), *Workforce readiness: Competencies and assessment* (pp. 151-174). Hillsdale, NJ: Erlbaum.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Cognitive skills and their acquisition* (pp. 141-189). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*, 997-1003.

Cooke, N. J., Rivera, K., Shope, S.M., & Caukwell, S. (1999). A synthetic task environment for team cognition research. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting,* 303-307.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. (2000). Measuring team knowledge. *Human Factors, 42*, 151-173.

Cooke, N. J., Shope, S.M., & Rivera, K. (2000). Control of an uninhabited air vehicle: A synthetic task environment for teams. *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting.*

Cooke, N. J., Stout, R., & Salas, E. (1997) Expanding the measurement of situation awareness through cognitive engineering methods, *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting,* 215-219.

Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review, 80(2)*: 97-125.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., & Nikolic, D. & Manning, C. A. (1998). Situation awareness as a predictor of performance in en route air traffic controllers. *Air Traffic Control Quarterly.*

Glaser, R. & Chi, M. T. H. (1988). Overview. In M.T.H. Chi, R. Glaser, and M.J. Farr (Eds.), The *Nature of Expertise* (xv-xxviii). Hillsdale, NJ: Erlbaum.

Gugerty, L., DeBoom, D., Walker, R., & Burns, J. (1999). Developing a simulated uninhabited aerial vehicle (UAV) task based on cognitive task analysis: Task analysis results and preliminary simulator data. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 86-90). Santa Monica, CA: Human Factors and Ergonomics Society.

Hinsz, V. B. (1999). Group decision making with responses of a quantitative nature: The theory of social decision schemes for quantities. *Organizational Behavior and Human Decision Processes, 80(1)*: 28-49.

Howell, W. C., & Cooke, N. J. (1989). Training the human information processor: A look at cognitive models. In I. Goldstein (Ed.), Training *and Development in Work Organizations: Frontier Series of Industrial and Organizational Psychology, Volume 3, New* York: Jossey Bass, 121-182.

Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management, 20,* 403-437.

Langan-Fox, J., Code, S., & Langfield-Smith, K. (2000). Team mental models: Techniques, methods, and analytic approaches. *Human Factors, 42*: 242-271.

Martin, E., Lyon, D. R., & Schreiber, B. T. (1998). Designing synthetic tasks for human factors research: An application to uninhabited air vehicles. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 123-127). Santa Monica, CA: Human Factors and Ergonomics Society.

Orasanu, J. (1990). Shared mental models and crew decision making. (Tech. Rep. No. 46). Princeton, NJ: Princeton University, Cognitive Science Laboratory.

Salas, E. Cannon-Bowers, J.A., Church-Payne, S., & Smith-Jentsch, K. A. (1998). Teams and teamwork in the military. In C. Cronin (Ed.), *Military Psychology: An Introduction* (pp. 71-87). Needham Heights, MA: Simon & Schuster.

Salas, E. Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an understanding of team performance and training. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 3-29). Norwood, NJ: Ablex.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization.* Norwood, NJ: Ablex.

Stout, R., Cannon-Bowers, J. A., & Salas, E. (1996). The role of shared mental models in developing team situation awareness: Implications for training. *Training Research Journal, 2,* 85-116.

Wickens, C. D. (1998). Commonsense statistics. *Ergonomics in Design, 6,* 18-22.

ACKNOWLEDGEMENTS

This work was supported by AFOSR Grant No. F49620-98-1-0287. The authors are grateful to NMSU's Air Force ROTC for their contribution to this work as team participants and to Greg Bromgard, Amy Burgess, Steve Jackson, Skye Pazuchanics, Harry Pedersen, Rebecca Serros, Natahsa Sutherlin, and Bien van der Meulen for their help with setting up, conducting, and analyzing data for this study.

Table 1. Nine process behaviors and associated event triggers for missions 1-6, 8, and 9.

TRIGGER: BEGINNING OF MISSION
(1) In first 5 minutes of mission team makes planning statements.

TRIGGER: LVN-OAK OR FIRST ROZ BOX
(2) Prior to UAV in effective radius (within 5 miles of) of H-AREA or F-AREA or targets within first ROZ BOX, DEMPC communicates restrictions on H-AREA and/or F-AREA to AVO.

(3) Prior to UAV in effective radius (within 5 miles of) of H-AREA or F-AREA or targets within first ROZ BOX, AVO acknowledges the DEMPC's communication or requests the information.

(4) Prior to UAV in effective radius (within 5 miles of) of H-AREA or F-AREA or targets within first ROZ BOX, DEMPC communicates upcoming targets (H-AREA, F-AREA) to PLO.

(5) Prior to UAV in effective radius (within 5 miles of) of H-AREA or F-AREA or targets within first ROZ BOX, PLO acknowledges the DEMPC's communication or requests the information.

TRIGGER: AFTER KGM-FRT CALL-IN
(6) Within 5 minutes after call-in of new ROZ box (KGM-FRT) DEMPC communicates new ROZ (KGM-FRT) and new targets to AVO and PLO

TRIGGER: PRK-ASH OR SECOND ROZ BOX
(7) Prior to UAV in effective radius (within 5 miles of) of S-STE or MSTE or targets within second ROZ box DEMPC anticipates PLO's need and communicates the PRK-ASH targets (S-STE, MSTE) without PLO asking.

(8) While UAV within PRK-ASH ROZ box (e.g., 2.5 miles of PRK or ASH) or within second ROZ box AVO and PLO work together to maneuver UAV for photos (this should be evident in their communication).

TRIGGER: END OF MISSION
(9) Within 5 minutes after end of mission team assesses and discusses their performance.

Table 2. Knowledge measures used in this study and associated characteristics and results.

Knowledge Measure	**Knowledge Duration** *L = Long-term* *F =Fleeting*	**Knowledge Type** *A= Taskwork* *E = Teamwork*	**Method Innovations** *HT = Heterogeneous Teams* *HO = Holistic*	**Timing of Measure** *M = During Mission* *K = During Knowledge Session*	**Results** *A= Revealed Acquisition* *P = Predicted Performance*
Team Situation Models	F	A		M	A, P
Teamwork Questionnaire	L	E		K	A
Taskwork Questionnaire	L	A		K	
Taskwork Ratings	L	A	HT	K	P
Taskwork Consensus Ratings	L	A	HO	K	P

Table 3. Procedures during each of three sessions.

SESSION	PROCEDURE
Session 1	PowerPoint Training
	Skill Training
	Break
	Mission 1
	Knowledge Session 1
	Mission 2
	Mission 3
Session 2	Mission 4
	Knowledge Session 2
	Break
	Mission 5
	Mission 6
	Break
	Mission 7
	Knowledge Session 3
Session 3	Mission 8
	Mission 9
	Break
	Knowledge Session 4
	Mission 10
	Debriefing

Table 4. Mean team performance scores, team process scores, and team SM (situation model) measures across the 10 missions.

MISSION	Team Performance	Team Process	Team SM Query 1 & 3 accuracy	Team SM Query 1 & 3 similarity	Team SM Query 2 - who?	Team SM Query 2 - topic?
Mission 1	509.5	.735	.788	.854	.788	.583
Mission 2	735.3	.823	.868	.889	.792	.558
Mission 3	821.8	.832	.886	.881	.798	.536
Mission 4	885.9	.849	.940	.928	.843	.546
Mission 5	896.6	.859	.956	.970	.783	.458
Mission 6	908.2	.843	.983	.970	.758	.508
Mission 7	910.0	.864	.959	.937	.800	.500
Mission 8	805.2	.867	.887	.942	.704	.509
Mission 9	883.7	.778	.936	.959	.783	.625
Mission 10	881.3	.815	.943	.935	.905	.619

Table 5. Correlations between long-term knowledge measures taken at Session 1 and performance across the ten missions. Pearson correlations are based on data from eleven teams (df = 9) except for Missions 8 and 9 (10 teams) and Mission 10 (9 teams). With 9 degrees of freedom *r* of .52 is significant at the *p* = .10 level. (* p < .10)

MISSION	Teamwork Questionnaire		Taskwork Questionnaire		Taskwork Ratings				Taskwork Consensus Ratings
	Accuracy	*Similarity*	*Accuracy*	*Similarity*	*Accuracy*	*Similarity*	*Role Accuracy*	*IPK Accuracy*	*Accuracy*
Mission 1	-.127	-.174	.143	-.068	.535*	.578*	.186	.232	.377
Mission 2	-.379	.041	-.08	-.380	.839*	.748*	.354	.582*	.252
Mission 3	-.122	.05	-.324	-.532	.769*	.684*	-.048	.605*	.502
Mission 4	.07	-.103	-.321	-.473	.770*	.738*	.004	.613*	.549*
Mission 5	-.382	.084	.279	.001	.485	.505	.443	.368	-.162
Mission 6	-.410	-.017	.548*	.109	.329	.274	.265	.168	-.115
Mission 7	.196	-.053	.232	.105	.085	.094	-.439	-.024	.551*
Mission 8	.037	.419	-.190	-.366	.382	.431	-.116	.555*	.078
Mission 9	-.366	.263	-.215	-.419	.669*	.600*	.146	.557*	.210
Mission 10	-.178	.408	-.271	-.490	.725*	.640*	.045	.677*	.302

Table 6. Classification of Knowledge Session 1 rating pairs on the basis of mapping individual to team consensus ratings. Asterisks indicate that strategy for that team occurred more than expected by chance.

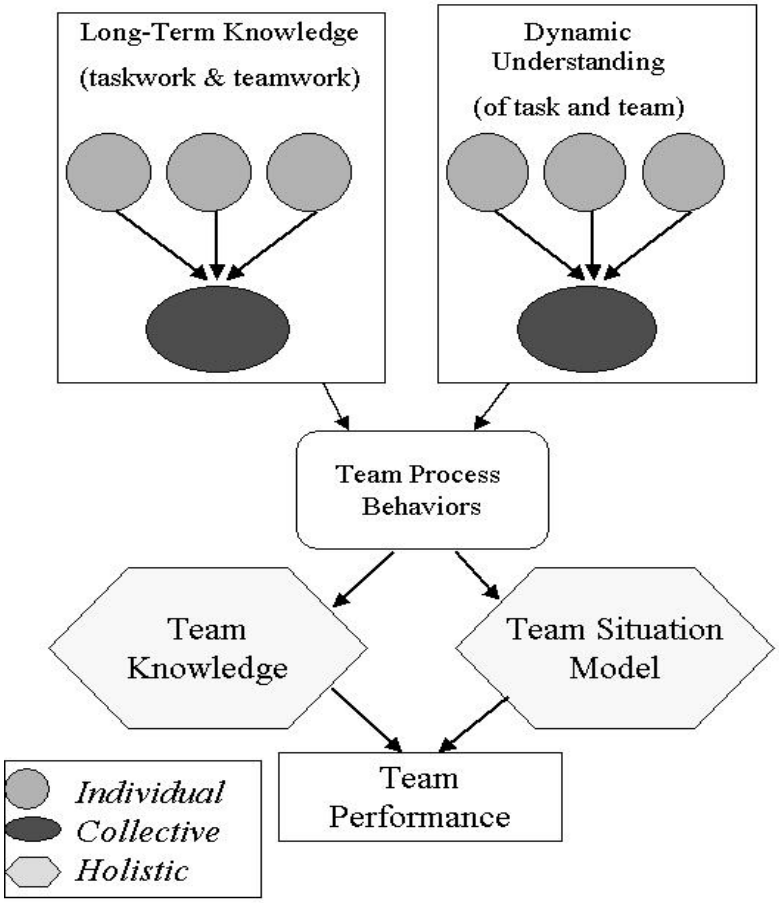
TEAM	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5
1	*13	3	*15	9	*15
2	8	*16	*13	4	*14
3	1	*21	*14	*14	5
4	11	8	11	10	*15
5	7	*16	*20	4	8
6	9	*15	*13	4	*14
7	10	10	10	8	*17
8	*15	*16	*12	3	9
9	6	*12	6	6	*25
10	6	*12	*16	6	*15
11	5	*19	*16	4	11
TOTAL	**91**	***148**	***146**	**72**	***148**

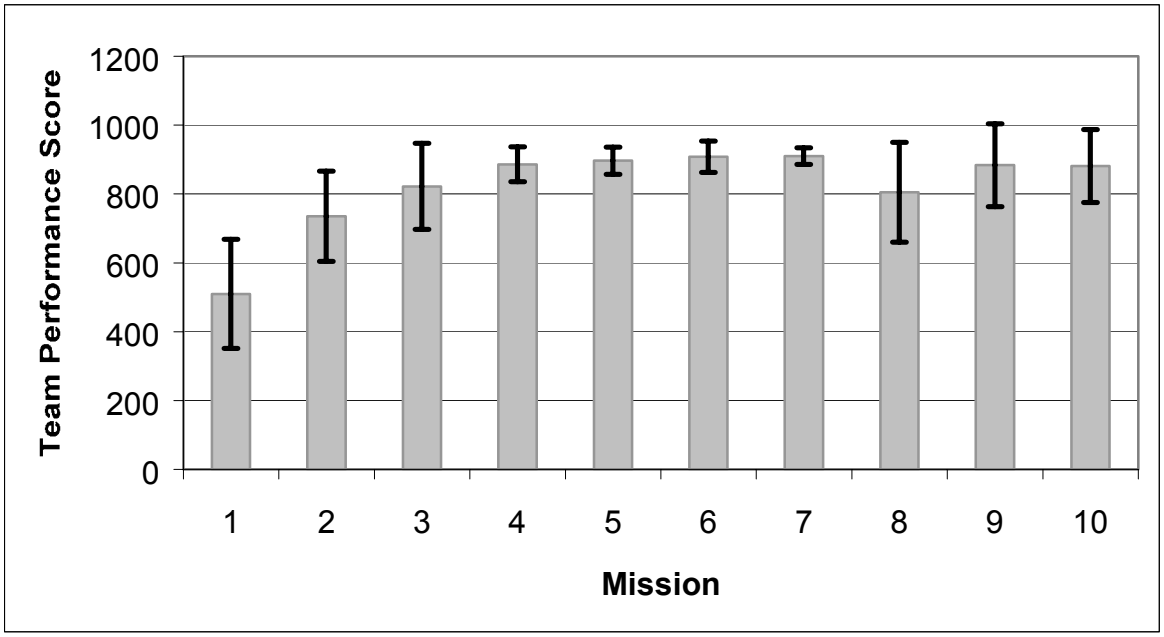
FIGURE CAPTIONS

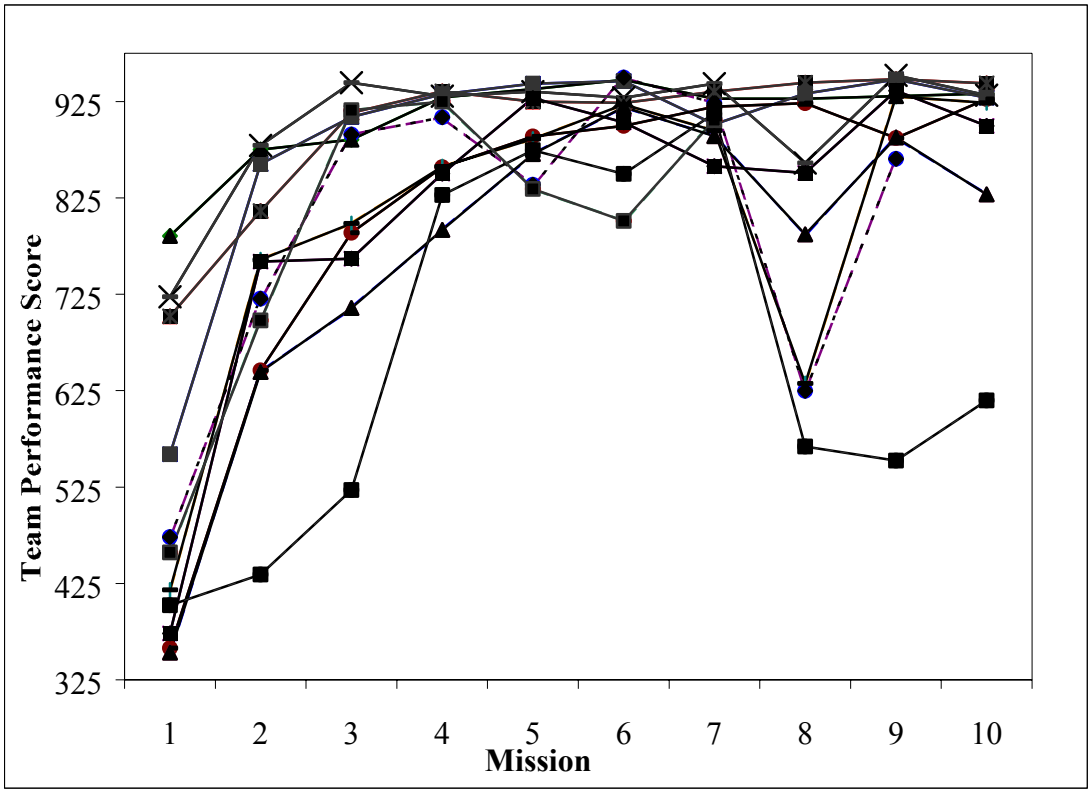
Figure 1. Framework for team knowledge.

Figure 2. Mean composite performance scores and standard deviations (represented by brackets) across teams for each mission.

Figure 3. Composite performance scores for each of 11 teams across each of the 10 missions. Missions 7 and 10 were associated with a scenario different from the other missions.


Long-Term Knowledge
(taskwork & teamwork)
Dynamic
Understanding
(of task and team)
Team Process
Behaviors
Team
Knowledge
Team Situation
Model
Team
Performance
Individual
Collective
Holistic


Team Performance Score
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
Mission


Team Performance Score
925
825
725
625
525
425
325
1
2
3
4
5
6
7
8
9
10
Mission