MILITARY PSYCHOLOGY, 2003, *15*(1), 77–96

The Importance of Crew Resource Management Behaviors in Mission Performance: Implications for Training Evaluation

Robert T. Nullmeyer
Aircrew Training Research Division
Air Force Research Laboratory
Mesa, Arizona

V. Alan Spiker
Anacapa Sciences
Santa Barbara, California

Cockpit/crew resource management (CRM) training within the military has grown rapidly despite the paucity of empirical data linking CRM to mission performance. CRM training objectives (and course content) are often too vague to allow meaningful training evaluation within the context of traditional transfer-of-training paradigms. A multimeasure methodology that exploits all sources of archival and observational data within a training organization has the potential to advance training evaluation, particularly for crew-based skills such as CRM. This article discusses a variety of CRM data sources and presents findings using 2 of these sources: instructor comments in student training folders and over-the-shoulder observations of crews in tactical simulators. Instructor comments revealed that CRM problems early in training most frequently involve decision making and communication among crew members. Over-the-shoulder observations of experienced crews showed high correlations between independent ratings of CRM proficiency and mission performance. The most effective crews exhibited such characteristic CRM behaviors as the presence of a single leader and willingness to change plans based on changing mission situations. The article closes by describing how these study data can be used to restructure CRM training into a set of behavior-based objectives that will enable meaningful evaluation of its effectiveness in improving the performance levels of all student crews.

Requests for reprints should be sent to Robert T. Nullmeyer, Air Force Research Laboratory, 6030 South Kent Street, Mesa, AZ 85212–6061. E-mail: robert.nullmeyer@williams.af.mil

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 DEC 2003	**Journal Article**	**01-01-2001 to 30-11-2003**

4. TITLE AND SUBTITLE
The Importance of Crew Resource Management Behaviors in Mission Performance: Implications for Training Evaluation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62205F

6. AUTHOR(S)
Robert Nullmeyer; V. Spiker

5d. PROJECT NUMBER
1123

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory,Aircrew Training Research Division,6030 South Kent Street,Mesa,AZ,85212-6061

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL/HEA

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory, Warfighter Training Research Division, 6030 South Kent Street, Mesa, AZ, 85212-6061

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RH; AFRL/RHA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RH-AZ-JA-2003-0002

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Published in Military Psychology, 2003, 15(1), 77-96

14. ABSTRACT
Cockpit/crew resource management (CRM) training within the military has grown rapidly despite the paucity of empirical data linking CRM to mission performance. CRM training objectives (and course content) are often too vague to allow meaningful training evaluation within the context of traditional transfer-of-training paradigms. A multimeasure methodology that exploits all sources of archival and observational data within a training organization has the potential to advance training evaluation, particularly for crew-based skills such as CRM. This article discusses a variety of CRM data sources and presents findings using two of these sources: instructor comments in student training folders and over-the-shoulder observations of crews in tactical simulators. Instructor comments revealed that CRM problems early in training most frequently involve decision making and communication among crew members. Over-the-shoulder observations of experienced crews showed high correlations between independent ratings of CRM proficiency and mission performance. The most effective crews exhibited such characteristic CRM behaviors as the presence of a single leader and willingness to change plans based on changing mission situations. The article closes by describing how these study data can be used to restructure CRM training into a set of behavior-based objectives that will enable meaningful evaluation of its effectiveness in improving the performance levels of all student crews.

15. SUBJECT TERMS
Crew resource management; Mission performance; Training evaluation; Cockpit resource management; Transfer of training; CRM; Decision making; Crewmember communication; Crew-based skills; Flight simulators

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Public Release**	**20**	

Human error is frequently linked with aviation accidents and incidents (Kayten, 1993). In fact, Helmreich and Foushee (1993) reported that aircrew actions were causal factors in more than 70% of hull-loss accidents in the worldwide commercial jet fleet from 1959 through 1989. In Ruffell Smith's (1979) landmark simulator study on cockpit workload, the factors that most differentiated effective crews from weaker ones were leadership, decision making, and resource management rather than more technically oriented skills. In response to findings such as these, the National Aeronautics and Space Administration (NASA) sponsored a workshop, *Resource Management on the Flightdeck,* in 1979 (Cooper, White, & Lauber, 1980), which is commonly viewed as the origin of formal cockpit/crew resource management (CRM) training. NASA and the Military Airlift Command cosponsored a follow-up conference on CRM training several years later (Orlady & Foushee, 1987). This workshop marked the expansion of CRM training into the military services. Each of the military services quickly added CRM training to selected programs, and within a few years, CRM instruction became mandatory for all military aviators.

This rapid growth of CRM training throughout aviation occurred despite surprisingly little empirical evidence linking this training to improved mission performance. The CRM validation data that do exist are dominated by trainee judgments about its value (e.g., Ilgen, 1999; Salas, Fowlkes, Stout, Milanovich, & Prince, 1999). Despite its short history, five distinct generations of CRM training can already be documented in the commercial airlines (Helmreich, Merritt, & Wilhelm, 1999), with each generation representing a substantial shift in training philosophy and content.

The military's approach to measuring the effectiveness of training interventions—be they a revised program of instruction, improved courseware, or new training device—has traditionally followed Kirkpatrick's (1996) four-stage model, depicted in the middle of Figure 1 (Bell & Waag, 1998; Salas et al., 1999). Although effectiveness is ultimately equated with the "contribution of training to the required availability of combat power" (Stage 4; Bell & Waag, 1998, p. 234), the vast majority of effort is focused on the earlier stages. These are the trainees' perceived value of the training, the degree to which the to-be-trained knowledge–skills–attitudes (KSAs) are actually learned, and to a much lesser extent, the availability of targeted KSAs for use on the job. From a research standpoint, Stage 3 is the defining feature of effectiveness, as demonstrated by a positive transfer to the job (combat) environment. Corresponding to each stage is a traditional methodology for evaluating training, as shown in the left part of Figure 1. These include a survey of trainees' attitudes toward their training, a pre- and postexperimental comparison of trainees' performance on the KSAs, a full transfer-of-training study conducted in the job environment, and a cost-effectiveness analysis that demonstrates the desired organizational productivity or performance impact (e.g., combat readiness) at an acceptable cost.

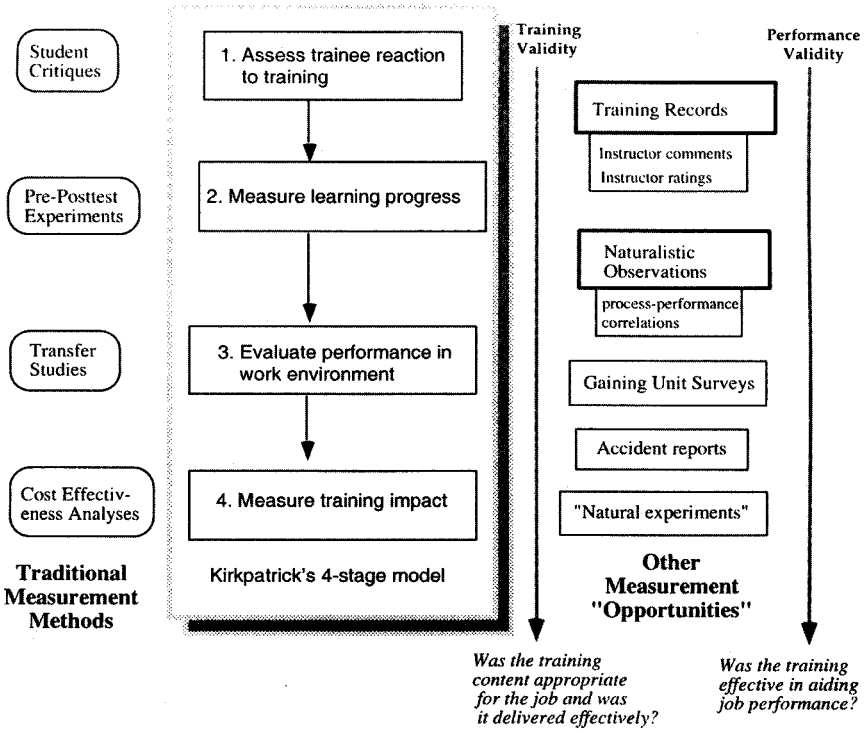

FIGURE 1 Framework for measuring training effectiveness.

Though influential, the model's limitations as a comprehensive theory of effectiveness measurement have been noted by a number of researchers. For example, Bell and Waag (1998) acknowledged the "brute force" aspect of the model's sequential stages and cited the need to distinguish between training processes and performance, particularly in Stage 2. Salas et al. (1999) used the model's hierarchy primarily as a theoretical "driver" for collecting data within a multimeasurement framework on trainees' reactions, attitudes, knowledge, and behavior, with no sequential priorities imposed on the measures.

Bell and Waag (1998) and Thurman and Dunlap (1999) reported a paucity of transfer studies corresponding to Stage 3 of Kirkpatrick's (1996) model. This dearth is tied to cost, methodological and operational constraints associated with testing newly trained (and now valuable) personnel in actual or near-realistic mission scenarios, soliciting and preserving a suitable control group, and obtaining the neces-

sary personnel stability within a turbulent environment (Thurman & Dunlap, 1999). Barriers to effective transfer are many, with perhaps the most significant being no clear definition of what is being trained. The Kirkpatrick model was promulgated in an era when procedurally oriented KSAs from prescribed task to training lists (TTLs) formed the content for training (Bills & Wood, 1999). With the proliferation of simulation devices geared toward facilitating cognitive-based processes such as decision making and situation awareness, the task-specific basis of training content has been altered. This has been particularly true in CRM, where—depending on the researcher or the tactical domain—it has been referred to as a training program (Ilgen, 1999), instructional strategy (Salas et al., 1999), tacit knowledge, metaskill, or enterprise (Bills & Wood, 1999).

In discussing training program evaluation, Goldstein (1987) distinguished between gauging the "training validity" of the intervention, that is, determining if training content is appropriate and effectively delivered, and gauging its "performance validity," that is, the extent to which the intervention facilitates subsequent job performance. As denoted by the arrows in the right-hand portion of Figure 1, these two aspects of training effectiveness can be assessed in parallel, and they extend across all stages of Kirkpatrick's (1996) model.

Given the limitations of the stage model discussed previously and the attendant difficulties in conducting transfer-of-training experiments, we believe that a more viable framework for measuring training effectiveness will emphasize capitalizing on multiple measurement opportunities, much as is done in the program evaluation area (Cook & Campbell, 1979). Considering the typical military training environment, there are a number of alternative data sources that bear directly on the effectiveness of a unit's training. Five of these opportunities are indicated in the right-hand portion of Figure 1. Although these are ordered according to their approximate placement within Kirkpatrick's (1996) stages, each method can apply to either training validity or program validity; consequently, we have not attempted to specify formal links with any of the stages.

Naturalistic observations of military aviators during simulator missions have revealed strong empirical links between the quality of crew interactions (process) and mission performance (outcome). Povenmire, Rockway, Buenecke, and Patton (1989) observed seven B-52 crews execute a tactically realistic scenario in a high-fidelity simulator and reported a statistically significant rank order correlation between mission performance and CRM ($r = .83$). Practicing inquiry and advocacy, avoiding distractions, distributing workload, and resolving conflicts emerged as factors that were significantly correlated with overall crew coordination.

Thornton, Kaempf, Zeller, and McAnulty (1992) observed 19 pairs of aviators as they flew a combat-oriented mission in an advanced UH-60 Black Hawk simulator. Aircrew coordination was defined in terms of the rate, pattern, content, and quality of interactions along 13 functional categories (inquiry, command, declarative, etc.). Mission effectiveness was defined in terms of navigation accuracy,

threat avoidance, and performance of a nonprecision approach. Patterns and types of communication were related to outcome-based indexes of mission performance, but rate of communication was not.

Brannick, Prince, Prince, and Salas (1995) also reported a strong empirical relation between team coordination and performance in a Navy study in which 52 two-person crews flew nontactical scenarios in a low-fidelity, tabletop T-44 flight trainer. Six team coordination dimensions were rated: assertiveness, decision making, adaptability, situational awareness, leadership, and communication. A rating scale ranging from 1 (*unacceptable*) to 5 (*excellent*) was developed for each dimension, and behaviors representing the performance expected were provided for the scale points. For example, assertive behaviors associated with higher ratings included questioning some directions from air traffic control and admitting confusion about an assigned altitude. All six process dimensions were positively correlated with performance, with correlations ranging from .43 to .69.

The remaining three methods in Figure 1 illustrate how effectively training has prepared students for performing their designated roles in their operational units. These measures include reviewing the surveys completed by training unit supervisors concerning how well trainees could support the unit without additional training, tracking and trending accident reports to identify links with common sets of prior training experiences, and examining the conduct of "natural experiments" that arise due to operational circumstances. An example of natural experiments involved a comparison of mission readiness during the Gulf War for groups of aviators who either did or did not receive full-mission simulator training (the simulator was unavailable for technical reasons for a cohort of aviators) prior to their arrival in the Middle East (Rakip, Kelly, Appler, & Riley, 1993). That simulator-trained MH-53J Pave Low operators were more able to conduct missions immediately on arrival in a country compared with their nonsimulator-trained counterparts was considered to be evidence supporting the effectiveness of the MH-53J flight simulator.

Ultimately, CRM training effectiveness evaluation must address the degree to which actual training needs are satisfied. Competing definitions and poorly defined training objectives have undoubtedly contributed to the current dearth of empirical CRM training studies. The studies reported here were designed to define CRM in concrete, observable terms for MC-130P crew training at the 58th Special Operations Wing, Kirtland Air Force Base, New Mexico. Our objectives were both to establish the content of CRM training and to develop measures for subsequent training evaluation.

The top two boxes on the right portion of Figure 1 highlight the methods described in this article. Although each is described in more detail in the following sections, analysis of training records and naturalistic observations were methods of choice because they shed light on the relevance of existing CRM course content for MC-130P crews and the relation of putative CRM processes to mission performance. By content analyzing the instructors' comments in student grade folders

for simulator and flight-line missions, we may ascertain the CRM problems experienced by students in training. By augmenting these analyses with in-depth, naturalistic observations of crews during simulator and flight-line training, we may construct a better picture of which CRM behaviors contribute to successful mission performance and, of these, which are presently taught during CRM training.

STUDY 1: USING STUDENT GRADE FOLDERS TO ASSESS CRM PROFICIENCY

Overview

Analysis of archival training records, such as student grade folders, provides at least three advantages as a source of training effectiveness data. First, training records are routinely maintained by training wings and do not require additional effort to collect. Second, they are populated by input from instructors who are experts in that student's crew position and who have acquired considerable familiarity with each student's capabilities. Third, they are often quantitative in nature and hence amenable to statistical analysis.

MC-130P Student Grade Folders

Grade folders are maintained for each MC-130P student as he or she completes mission qualification (MQ) training. For each academic, simulator, and flight-line training session, the instructor assigns a letter grade (P = *proficiency advance,* E = *exceptional,* S = *satisfactory,* T = *needs training,* U = *unsatisfactory,* I = *incomplete*) in the student's Form 15, Aircrew Training Record. To be a useful source of proficiency data, a measure must vary across students; otherwise, one cannot infer the impact of program variables or training interventions. However, in our analysis of MC-130P student records, more than 98% of student grades in the Form 15 were assigned an "S."

MQ training progresses through seven simulator training events: a conversion mission, two day-tactical missions, and four night-tactical missions. Each block of simulator training is followed by flight-line instruction of corresponding complexity. To keep track of this progression, MC-130P instructors fill out a Form 14, Aircrew Training Progress Record, after each simulator and flight-line mission. This preprinted form provides a set of required proficiency levels (RPLs) for the training events associated with that mission profile. Events are task based, such as airdrop checklist, simulated engine failure, night vision device operations, and so forth. Performance and knowledge are graded on a 4-point scale, ranging from 1 (*extremely limited*) to 4 (*highly proficient*). As the student progresses through training, the RPLs for the events in each training profile become more stringent.

Instructors cross off RPLs on the preprinted form as each training event is accomplished. If a student exceeds or fails to meet an RPL, the instructor must write in the actual level of performance or knowledge demonstrated; however, instructors rarely note deviations from the RPL because they are busy and do not wish to have student deficiencies noted in the permanent record. Consequently, analysis of the RPL data, such as aggregating the number of sub-RPL events, will not produce useful measures of proficiency because most entries are the unannotated RPLs.

In addition to grades, instructors write comments following each simulator and flight session on Form 13, Training Comments Record. The comments are unstructured and are not necessarily tied to the required items covered in the Form 14. These comments provide information that is potentially a rich source of proficiency data as instructors are free to express their reservations regarding a student, knowing that their remarks are not reflected in the student's recorded grade. Instructors may also laud exemplary performance and can go over the comments with the student after a training mission, using it as a teaching or debriefing aid.

In analyses of MC-130P crew training records for each position, we determined that instructors make extensive comments that, when aggregated across missions, can be reliably classified into positive and negative cases. Moreover, the comments can be sorted into functional categories characteristic of each crew position (e.g., crew coordination, equipment knowledge). This analysis revealed that instructor comments are quite specific (e.g., "missed several radio calls," "must keep checklist flowing to ensure proper crew responses," "need more positive continuous guidance to pilot") and yield valuable insights concerning areas where student proficiency is strong or weak. To the extent that these comments are recorded routinely and comprehensively, they can be content analyzed, aggregated, and quantified to yield data-based assessments of student proficiency and, ultimately, training effectiveness.

Method

Based on the preliminary analyses noted previously, we used the instructor comments from student grade folders to assess the effectiveness of the wing's current CRM training. We reviewed a representative sample of 20 records from the five crew specialties in the MC-130P: pilots, navigators, flight engineer (FE), communication systems operator (CSO), and loadmaster (LM). We then enlisted the help of two subject matter experts (SMEs) to review the records according to the methodology outlined in the following paragraph. Both were experienced instructors, one in airborne command and control aircraft and the other in special operations, fixed-wing aircraft.

Working independently, the SMEs reviewed each training record and highlighted all instructor comments relevant to CRM. These comments were then paraphrased and transcribed onto a four-page, structured Training Record Evaluation

Worksheet. The worksheet was organized around the six CRM areas currently covered in Air Force Instruction 11–290 (*Cockpit/Crew Resource,* 1998). These CRM areas are mission planning and debrief, task management, situation awareness, crew coordination, communication, and risk management decision making. We added a seventh area, tactics employment, to address the combat-intensive operations required of this aircraft. Each paraphrased comment was placed in the relevant CRM category and then rated on a 5-point scale, ranging from 1 (*significantly below expectations*) to 5 (*exceptional*), with a midpoint of 3 (*level expected for this level of training*). Once comments were transcribed and rated, a summary rating for that CRM category was assigned. Finally, an overall proficiency rating, again on a 1 to 5 scale, was assigned for each student.

Results

Interrater reliability. The first goal of the analysis was to determine if the two SMEs were consistent in their rating assignments. The correlation between the two sets of 20 overall proficiency ratings was .81, within the recommended range for acceptable interrater reliability (Cronbach, 1990). Looking at the ratings themselves, we found that the two raters produced identical ratings for 16 of the students and differed by only one point for the other four students. We thus conclude that SMEs can reliably assign proficiency ratings based on the comments that instructors place in student grade folders.

Inferring student proficiency. Having established the reliability of the rating process, we had one of the SMEs extend his review to include all MC-130P student records from 1998, a total of 87 records. We then examined the absolute values of the ratings to determine if there was sufficient variation across students to assess CRM training. Of the 87 records, more than one third received a rating other than 3, a much higher percentage than was evident in the instructor grades on the Form 14s.

Even greater sensitivity is seen when the comment-based proficiency ratings are broken down by CRM category, as shown in Figure 2. Computing the average rating variability within each category, we determined that a scale difference of .14 can be considered statistically meaningful (Hays, 1973). On that basis, we see that mission preparation and crew coordination received substantially above-average ratings (3.2–3.3), with decision making and communications (2.8–2.9) significantly below average.

Qualitative analysis. Having identified the CRM categories that stand out statistically, we can examine the associated instructor comments to pinpoint areas where present CRM behaviors are strong and weak. In performing this analysis, it should be noted that such comments have two aspects: evaluative and directive. The evaluative component is used to gauge student proficiency in the commented area and is typically represented by an adjective, such as "good" mission planning,

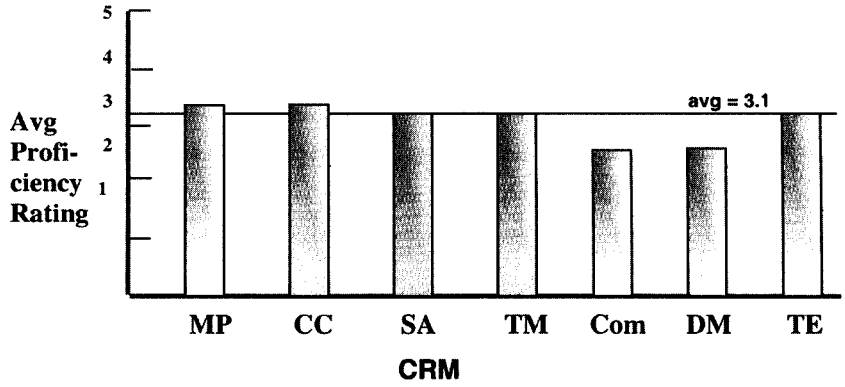

FIGURE 2 Average student proficiency by CRM category. CRM = cockpit/crew resource management; MP = mission planning and debrief; CC = crew coordination; SA = situation awareness; TM = task management; COM = communication; DM = risk management decision making; TE = tactics employment.

"excellent" mission briefing, or "weak" situation awareness. The directive aspects of each comment let us extract the specific crew behaviors that were either deficient or laudable. By directive, we mean such comments as "slow to prepare brief," "needs to think further ahead of the aircraft," and "missed radio calls." These comments are usually given to the student as verbal feedback during the training session to promote immediate improvement or to reinforce some essential skill. Over the long term, the content of these comments can be collected, analyzed, and folded back into an improved training curriculum as a set of target behaviors.

To illustrate, we reviewed the compiled set of instructor comments in mission preparation and crew coordination to ascertain why these categories stood out as positive instances of CRM. For the most part, we saw that instructors were complimenting students on such aspects as "thorough" planning, "concise" briefings, and "good backing up crew members." These areas are emphasized in the present training and seem to have been internalized by the students. Yet there were also negative comments that indicated areas in need of improvement, such as navigation and leadership. Examples here include the need for providing greater annotation of significant terrain features on maps, discussing more obstacles in the low-level brief, and taking firmer control of the crew.

Turning to the weaker CRM areas, decision making exhibited a wide assortment of deficiencies that primarily involved pilots and navigators. A major deficiency entailed slow reactions to conditions requiring more rapid judgment, such as initiating emergency procedures, responding to loss of engine, turning to final approach, joining up during aerial refueling, correcting the flight profile, and breaking off formation during the onset of instrument meteorological conditions. A host of communication problem areas were also exposed for all crew members.

These included missing air traffic control calls (pilot), weakness in procedural terminology (LM), as well as the need to provide more guidance to pilot (navigator), break in when necessary (CSO), and be more assertive (FE).

Conclusions

These analyses indicate that training records, specifically the instructor comments, can serve as a reliable and efficient source of proficiency data from which training effectiveness may be inferred. Although our analysis was slanted toward CRM, other aspects of training (e.g., procedures, tactics, and equipment operation) could be scrutinized as well. From our experience, we believe any training organization where instructor comments are archived could employ the quantitative and qualitative analyses described previously to create a database for tracking and trending student proficiency across designated areas. This would provide baseline data from which the status of a training program could be discerned and could feed periodic reports to decision makers and training managers, thereby more effectively integrating training data with the management of wing operations.

STUDY 2: NATURALISTIC OBSERVATIONS OF CREW CRM BEHAVIORS

Overview

Fully qualified MC-130P crews from operational squadrons attend simulator refresher training annually. The fifth and final day of this training involves planning for, and then executing, a demanding tactical mission in a high-fidelity simulator. This training allows observation of crew behaviors throughout the entire mission sequence, from mission tasking to postmission debriefing. The training scenario includes less-than-perfect information about the tactical environment, irrelevant communications, and unforeseen events that, like the real world, occur in tactical missions.

The first mission execution event is night, low-level flight. Crews fly at low altitudes in response to potential enemy threats, with pilots using night vision goggles. The next mission phase requires in-flight air refueling of multiple helicopters within prescribed time, course, and altitude constraints. The third mission phase is an airdrop delivery of personnel in a potentially hostile area. The final phase, infiltration/exfiltration (infil/exfil), is to infiltrate a potentially hostile area, land at an unsecured airfield, pick up personnel, and return to friendly territory. Planning and executing these four mission phases requires substantial collaboration among crew members. Observations of CRM and mission effectiveness were organized around mission planning and these four mission execution phases.

Method

Participants. Thirteen experienced MC-130P aircrews were observed during their fifth day of annual simulator refresher training. Two crews were unable to fly the mission due to simulator malfunctions. Thus, 11 crews (66 crew members) were included in the final analysis. Individual crew member experience levels averaged 3,056 total flying hours and 1,286 MC-130P hours. The typical makeup of an MC-130P crew for refresher training consists of an aircraft commander (AC), a copilot (CP), two navigators, one FE, and a CSO.

Data collection instruments. The Team-Mission Observation Tool (T-MOT) was developed to aid in recording behaviors that fell into five categories that had been identified by SMEs as CRM areas in which MC-130P crews exhibited considerable variability. Function allocation referred to the division of crew member responsibilities to avoid redundant tasking and task overload. Tactics employment dealt with avoiding or minimizing exposure to threats and coordination of multiple mission objectives. Situational awareness reflected maintenance of an accurate mental picture of mission events and objectives as they unfold over time and space. Command–control–communication addressed activities involving external parties in the mission as well as communication within the crew. Time management focused on the ability of the crew to manage limited time resources.

These CRM categories were rated using 5-point scales, from 1 (*lowest*) to 5 (*highest*), with a 3 rating reflecting basic compliance with minimum command requirements. CRM behaviors were rated during five discrete mission phases: mission planning, low level, air refueling, airdrop, and infil/exfil. Activities and behaviors that seemed unusually strong or weak were also documented.

The Team-Mission Performance Tool (T-MPT) was designed to structure crew mission performance ratings in each mission phase. This instrument provided several 5-point, behaviorally anchored rating scales (BARS) for a second researcher to rate the quality of individual- and team-generated mission products developed during mission planning and also to rate demonstrated crew performance within each mission phase.

Procedure. Over-the-shoulder observations were the primary source of CRM and mission performance data, with one researcher collecting CRM data and the other, performance data. A highly trained, former MC-130P navigator recorded CRM data. During mission execution, this SME-researcher observed and monitored crews from an intercom station located outside the MC-130P Weapon System Trainer (WST). The SME-researcher was situated in front of four instructor–operator screens that displayed instructor input from inside the WST.

Performance data were collected similarly. The second researcher observed crew performance and recorded observations during the five mission segments.

Her observations and notes during planning captured such items as the number of briefings each crew gave, the content of the briefings, who performed the briefings, and the number of charts created. During mission execution, this researcher was purposely located away from the CRM researcher to maintain independence of ratings. From her location, she monitored crew communications, flight path, and threat laydown. Various Instructor Operator Station (IOS) pages were selected and printed out. At the conclusion of the simulator mission, this researcher collected all products (flight plans, charts, checklists, etc.) that the crew created for the mission and used them, along with the other data collected, to complete the T-MPT.

Results

Statistical considerations. Because all tests reported in this section use crew as the unit of analysis, our total sample of 11 and the resulting 9 degrees of freedom (i.e., for t tests, $df = N - 2 = 9$) seem rather small to achieve the statistical power required to establish a strong process–performance relation. Yet, the 11 crews in our sample constitute 26% of the 42 MC-130P aircrews that go through refresher training in an average year. Because we sampled a sizable proportion of the population, we were able to reduce our estimated variance of the sample mean by using a finite-population correction coefficient (Winkler & Hays, 1975). The correction coefficient decreases the observed sample variance by the square root of $(N - 1)/N - n$, where N is the population size, and n is the sample size. In our case, the reported t values in the next section have been increased by 20%, reflecting a 1.2 finite-population coefficient multiplier.

Due to the exploratory nature of many of our research questions and the need to perform a large number of statistical tests, we used a Bonferroni adjustment as a way to keep our overall, or experiment-wise, alpha level from exceeding the desired (nominal) level. The Bonferroni technique is a conservative, though effective, way to avoid inflating the alpha level (and hence the likelihood of a Type I error) caused by "snooping" through one's data to locate the largest effect (Harris, 1994). The adjustment is made by dividing the desired experiment-wise alpha level by the number of tests that are performed in a given cycle of testing. As shown at the bottom of Table 1, we employed a conservative nominal alpha level of .002 to control for the fact that we were reporting 31 statistical tests. Because all correlations were tested against a null hypothesis of zero, a critical t value of 4.19 is required to maintain an experiment-wise alpha level of .05. Using the finite-population correction described previously, our reported correlations have to reach a criterion of at least .74 to be significant, given the adjustment for multiple statistical tests.

Ratings of CRM proficiency and mission performance. The first question to address is whether CRM is an important predictor of mission performance. A strong, positive relation was found between overall CRM process and total mis-

TABLE 1
Correlations Between CRM Elements and Mission or Mission Phase Performance

	Whole Mission	*Mission Planning*	*Low Level*	*Airdrop*	*Air Refueling*	*Infil/Exfil*
Situational awareness	.76**	.48	.47	.59	.28	.65*
Tactics employment	.78**	.27	.06	.54	.81**	.55
Time management	.83**	.41	.36	.66*	.51	.64*
Function allocation	.75**	.22	.61*	.55	.55	.60
Communication	.08	.14	.09	.32	.37	.30

Note. CRM = cockpit/crew resource management; infil/exfil = infiltration/exfiltration.
*$p < .05$, unadjusted. **$p_{EW} < .05$, $p_{NOM} < .002$ Bonferroni adjustment assuming 31 tests; critical $r = .74$.

sion performance ratings ($r = .86$), $t(9) = 6.143$, $p < .001$. Figure 3 depicts the scatterplot between crew performance on the y axis and CRM process rating on the x axis. As is evident from the figure, the poorest performing crew did indeed have the lowest overall CRM process; the highest rated crews received the highest mission performance ratings. The intermediate values also behaved in a consistently ordered fashion.

Having established the fundamental relation between the CRM process and mission performance, our next step was to identify the CRM elements for which the relation is the strongest. The correlations that gauge the strength of the linear relation between each of the five CRM categories and overall mission performance are shown in the left-most data column in Table 1. Four elements were statistically

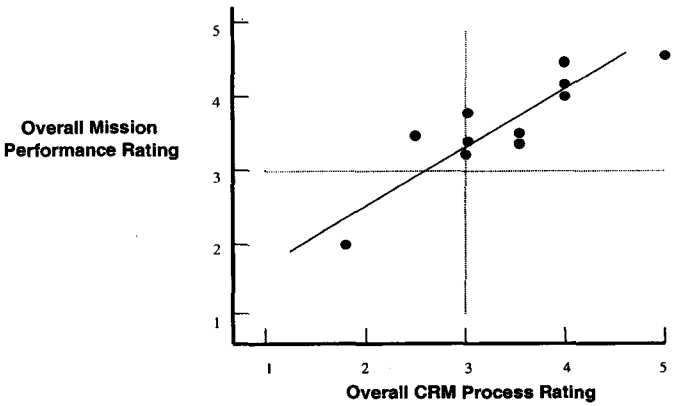

FIGURE 3 Correlation between overall CRM and mission performance ratings for MC-130P crews. CRM = cockpit/crew resource managment.

related to mission performance, with only command–control–communication failing to achieve significance. Interestingly, the relation between this element and mission performance was near zero for these experienced crews.

The remainder of Table 1 breaks down the analysis still further, into a 5 × 5 matrix of correlations between each phase-specific CRM subprocess rating and its corresponding phase-specific mission performance rating. All of the correlations were positive and some even quite large. However, using the stringent Bonferroni ($p < .002$) adjustment, only the correlation between the rating of tactics employment and mission performance during the air-refueling phase proved significant, $t(9) = 4.97, p < .001$. There is, however, some evidence that CRM subprocesses exhibited differential effects across mission phases using the traditional criterion of $p < .05$. SMEs uniformly indicated that this less conservative criterion was more useful for identifying CRM areas to emphasize in training. As we examine each column, we first see that, although no CRM element met either criterion for mission planning, the most strongly associated element is situational awareness. Function allocation is strongly associated with low-level flight performance ratings. The CRM element most strongly associated with airdrop performance is time management, although other elements also exhibited high positive correlations. Air refueling performance was most highly predicted by tactics employment ratings. Finally, infil/exfil performance had fairly large correlations with all of the CRM subprocesses save command–control–communication. Indeed, as can be seen in the bottom row of Table 1, communication did not have any correlations above .40 with any of the mission phases. This is consistent with the low correlation between the ratings of overall communication and overall mission performance.

Behaviors of effective crews. In addition to the quantitative ratings, extensive and detailed observations of CRM process and mission performance were collected and analyzed. In this section we summarize the CRM behaviors exhibited by the most effective crews. High situation awareness ratings were often associated with good mission performance. Behaviors that accompanied high ratings were as follows:

1. Giving greater consideration to the "big picture."
2. Viewing the crew as only part of the larger team and mission.
3. Raising extensive "what if" questions about main mission events, including input from the entire crew.
4. Accepting the need to change the plan, based on the evolving mission and changing situation, even if this caused some confusion to the crew.
5. Including explicit alternatives within the premission briefings, for example, no waiting on the air-refueling track if the helicopter is late, or cutting the jumper if someone gets hung.
6. Responding well to their own errors or changing conditions.

For tactics employment behaviors, effective crews included a separate threat briefing among their premission briefs. This briefing contained a detailed overview of the threat situation (especially "the shooters") and included preplanned crew responses to the threats depending on threat mode (i.e., search, acquisition, locked-on). Weaker crews omitted a threat briefing in their permission briefings or did not include variations in threat response based on specific threat modes. Other behaviors linked to high tactics-employment ratings included (a) asking if anyone had observed the threat visually, (b) asking about aircraft damage, (c) marking the latitude/longitude (lat/long) of the threat encounter, and (d) suggesting additional countermeasures even after successfully avoiding the threat (e.g., dispatching an F-16 to destroy target *X* in the event the crew was forced back into its vicinity later in the mission).

Time management is the fundamental feature of MC-130P missions and was highly correlated with mission performance. The primary behavior that characterized effective crews in this regard was their overt time awareness and monitoring throughout mission planning and execution. Exceptional crews would set times for briefings and mission tasks as well as ask about their progress toward these times. Within weaker crews, monitoring of time status was less apparent and overt, especially during planning. As a result, many weaker crews began briefings late or rushed through their briefings.

Function allocation was also highly correlated with mission performance. Individual crew member duties were overtly and explicitly designated in the most effective crews, based on crew member strengths rather than position. For example, an AC designated all communication responsibilities to the CSO. Although communication is the CSO's primary function, many crews opted for the navigator or CP or both to support these duties as well. This particular AC seemed to realize that mission complexity necessitated allocating this function solely to the CSO, so that navigators and CP could focus on other mission taskings. This explicitness was in contrast to the more implicit designation of duties in lower rated crews.

A few other process behaviors distinguished exceptional crews from weaker crews. However, these were not easily categorized into our five CRM categories. The following are the most salient "other" behaviors:

1. Exceptional crews were extremely focused on the mission with little (if any) socialization during planning; there was no chatter at all on the intercom during mission execution.
2. Exceptional crews tended to develop and use extremely aggressive plans.
3. Leadership emerged as a very powerful variable. Exceptional crews had a clear, single leader who worked to weave the crew together in all mission aspects including mission planning. Three of the four most effective crews had such a leader, resulting in highly integrated crews versus the fragmented dyads and triads seen in less effective crews.

Conclusions

Our quantitative analyses revealed substantial variability in both CRM and mission performance ratings, with very consistent links between the two. This suggests that CRM is indeed a strong predictor of mission performance, that some crews are better than others, and that these patterns are strong enough to be analyzed using traditional inferential statistics despite relatively small sample sizes. The qualitative observation data revealed consistent behavioral patterns that were unique to the crews that received exceptional mission performance ratings. None of these behaviors were included in existing CRM training, yet SMEs concurred with the value of the behaviors for mission performance. We concluded that CRM course content needed to be reviewed to determine if the most important areas are being covered. The wing concurred and established a working group to accomplish this review.

DISCUSSION

Rigid adherence to Kirkpatrick's (1996) four-stage model of training evaluation has resulted in lost opportunities to demonstrate effectiveness using methods other than the traditional, difficult-to-achieve, transfer-of-training paradigm. This is particularly the case for assessing the effectiveness of training skills such as CRM, whose underlying task basis is not well established. Following the lead recently taken by the Navy (Salas et al., 1999), we adopted a multimeasurement framework and employed a variety of converging data collection schemes to gauge CRM training effectiveness, including over-the-shoulder observations, training record analysis, accident reports, and gaining unit surveys. The two studies reported here support the value of capturing data from routine training to gain insight into the effectiveness of that training. The types of data presented—instructor comments in training records and over-the-shoulder observations of crew performance during simulator training—required no additional effort on the part of students, instructors, or training support personnel other than the researchers themselves. Such data traditionally have not been harnessed to improve training effectiveness.

Deficiencies in CRM-related skills during simulator and flight-line training were not evenly distributed across CRM areas for beginning MC-130P student crew members. Instead, decision making and communication showed unusually high concentrations of instructor comments in MQ training records, reflecting substandard student proficiency levels. Furthermore, several recurring themes emerged within these problem areas, none of which were specifically addressed in the student's preceding academic training. Addressing these recurring themes has the potential to enhance the ability of the training organization to meet its objectives, that is, reach Kirkpatrick's (1996) fourth stage, most notably by reducing the number of students who perform at less-than-criterion levels.

The naturalistic observations of MC-130P crew performance during annual refresher training provide insights about behaviors at the higher end of the experience spectrum. Perhaps the most notable finding was the strength of the relation between CRM proficiency and mission performance, enabling 75% of the observed variability in mission performance ratings to be accounted for by knowledge of rated CRM skill. There are at least two implications of this finding. First, with CRM accounting for such a large proportion of mission performance, a comprehensive training effectiveness evaluation must address this important category of behaviors. Second, because CRM ratings were based on observable crew behaviors, this strong correlation adds credibility to the behaviorally based approach to CRM training advocated by both the Navy (Prince & Salas, 1993) and Army (Leedom & Simon, 1995).

A second characteristic of the overall observational data was the rarity of substandard mission performance ratings, as only one of the 11 crews failed to meet at least minimum mission performance criteria. This is consistent with a historically low frequency of MC-130P accidents and incidents. Recognizing this limited range of mission performance scores for most MC-130P crews, we believe the most feasible goal in applying research results is to increase the proportion of crews exhibiting highly effective performance levels. As was the case with training record analysis findings, traditional CRM instruction does not reflect the key behaviors identified in the study. For example, despite a clear link between leadership and mission performance, this topic is not addressed at all in the wing's present CRM training.

The more detailed CRM–performance correlations reported in Table 1 suggest that some CRM elements are more central than others for effective performance during individual mission phases. For example, the diversity and complexity of activities involving multiple crew members during low-level flight create a high value for allocating crew member functions wisely during this phase. Time and accuracy are the two criteria used to measure performance during the airdrop mission phase. It is therefore not surprising that time management emerged as an important factor. Crews who excelled at this task called drop warnings and executed associated checklists in a timely fashion. The threat environment in this scenario rewarded crews who had focused on the tactical considerations associated with the planned air refueling track and maintained situational awareness during this phase of the mission. Infil/exfil required crews to integrate multiple information sources (e.g., intelligence reports on changing threats and the various parties on the ground), and like airdrop, the crew must reconfigure the airplane, run multiple checklists, and incorporate multiple crew member perspectives. Strong situational awareness and time management skills differentiated the most effective crews from the others.

The CRM training program for MC-130P crews generally corresponds to a second-generation CRM program as defined by Helmreich et al. (1999). The 12-hr

course includes eight seminar sessions covering CRM policy and regulations, command authority, aircrew communication, workload performance, available resources, situational awareness, decision making, and operating strategy. Each session applies to aviation in general but not specifically to the MC-130P aircraft, crew, or mission. Team-building exercises and aviation case studies are also included but, again, not tailored to MC-130P operations. Training objectives focus on academic knowledge (e.g., "define situational awareness," "describe the structured decision-making process," "list hazardous attitudes and their antidotes") rather than skills and behaviors.

Despite recent advances in our understanding of CRM and CRM training, Helmreich et al. (1999) observed that second-generation courses such as this one continue to be used. The CRM course for MC-130P crews was state of the art when implemented over a decade ago, but it has not been updated to reflect more recent CRM training concepts. As such, the course does not emphasize the specific skills and behaviors that characterized third-generation CRM, include the detailed analyses of aircraft-specific training requirements that characterize fourth-generation CRM, or address the error management aspects of fifth-generation CRM.

Subsequent to these studies, we reviewed Air Force CRM training regulations and found that this generic (second-generation) approach complied with a requirement to provide basic CRM instruction for all aviators. Historically, there may have been value in providing an introductory CRM course to all Air Force aviators. Over time, this approach has resulted in considerable duplication of instruction. Before MC-130P crews are given CRM training as part of the specialized MC-130P MQ course, CRM will have been introduced during undergraduate flying training and again as part of basic C-130 training, with introductory CRM training required at all three levels. Our analyses of aircrew behaviors revealed multiple shortfalls of generic CRM training for this particular student population at this stage of their training, especially given two previous exposures to similar or identical information.

The response from aviators to the results of these studies has been heartening. CRM training policy was recently changed to establish a "building-block" approach spanning an aviator's flying career (*Cockpit/Crew Resource,* 1998). Basic awareness, definitions, and general team dynamics are to be taught in undergraduate flying training. Follow-up CRM training for specific weapon systems will add specific aircraft and mission content, going beyond knowledge to include skills and behaviors. Data from our two studies are being applied in two ways to improve MC-130P CRM training. First, key behaviors from content analyses are being used to populate a new MC-130P CRM course with relevant content. Second, CRM areas that are highly correlated with mission outcome are reviewed by SMEs as areas that may warrant added emphasis. Substantial concurrence exists between the patterns of CRM–performance correlation noted in our data and the relative importance of CRM skills as perceived by SME instructors.

With respect to evaluating the effectiveness of the new CRM course, our findings have helped generate training objectives that are defined in terms of concrete, observable behaviors. Student training folder analyses revealed recurring CRM problems on the flight line (decision making and communication within crews) that are targeted in the new MC-130P CRM course. Observations of mission-qualified crews revealed those CRM behaviors consistently exhibited by the most effective crews. A goal of the new CRM course is to shape these behaviors in students before they get to the flight line. In both cases, we expect to see changes manifested in observable behaviors that can be captured and evaluated using the methods described in this article. Given the nature of the data collected so far, we are confident that well-defined training objectives, coupled with well-defined and observable measures of effectiveness, will support a credible training effectiveness assessment.

REFERENCES

Bell, H. H., & Waag, W. L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *The International Journal of Aviation Psychology, 8,* 223–242.

Bills, C. G., & Wood, M. E. (1999). New insight for training development of 21st century advanced warfighter training. In *Proceedings of the 1999 Interservice/Industry Training Simulation and Education Conference.* Arlington, VA: National Training Systems Association.

Brannick, M. T., Prince, A., Prince, C., & Salas, E. (1995). The measurement of team process. *Human Factors, 37,* 641–651.

Cockpit/crew resource management training program. (1998, July). (Suppl. No. AFI 11–290). Washington, DC: Headquarters U.S. Air Force.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation.* Chicago: Rand McNally.

Cooper, G. E., White, M. D., & Lauber, J. K. (Eds.). (1980). *Cockpit resource management training: Proceedings of a NASA/industry workshop* (Rep. No. NASA CP–2120). Moffett Field, CA: NASA–Ames Research Center.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper.

Goldstein, I. L. (1987). The relationship of training goals and training systems. In G. Salvendy (Ed.), *Handbook of human factors* (p. 964). New York: Wiley.

Harris, R. J. (1994). *ANOVA: An analysis of variance primer.* Itasca, IL: Peacock.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston.

Helmreich, R. L., & Foushee, H. C. (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 3–46). San Diego, CA: Academic.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *The International Journal of Aviation Psychology, 9,* 19–32.

Ilgen, D. R. (1999). Teams embedded in organizations: Some implications. *American Psychologist, 54,* 129–139.

Kayten, P. J. (1993). The accident investigator's perspective. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 283–314). San Diego, CA: Academic.

Kirkpatrick, D. (1996). Revisiting Kirkpatrick's four-level model. *Training and Development, 50*(1), 54–59.

Leedom, D. K., & Simon, R. (1995). Improving team coordination: A case for behavior based learning. *Military Psychology, 7,* 109–122.

Orlady, H. W., & Foushee, H. C. (Eds.). (1987). *Cockpit resource management training* (Rep. No. NASA CP 2455). Moffett Field, CA: NASA–Ames Research Center.

Povenmire, H. K., Rockway, M. R., Buenecke, J. L., & Patton, M. W. (1989). *Evaluation of measurement techniques for aircrew coordination and resource management skills* (Rep. No. UDR–TR–89–108). Chandler, AZ: Air Force Human Resources Laboratory, Operations Training Division, Williams Air Force Base.

Prince, C., & Salas, E. (1993). Training and research for teamwork in the military aircrew. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 337–366). San Diego, CA: Academic.

Rakip, R., Kelly, J., Appler, S., & Riley, P. (1993, November). The role of the MH-53J III E Pave Low weapon system training/mission rehearsal system (WST/MRS) in preparing students for Operation Desert Storm, and future operations. In *Proceedings of the 1993 Interservice/Industry Training Systems Conference* (pp. 432–438). Arlington, VA: National Training Systems Association.

Ruffell Smith, H. P. (1979). *A simulator study of the interaction of pilot workload with errors, vigilance, and decisions* (Rep. No. NASA TM–78482). Moffett Field, CA: NASA–Ames Research Center.

Salas, E., Fowlkes, J. E., Stout, R. J., Milanovich, D. M., & Prince, C. (1999). Does CRM training improve teamwork skills in the cockpit? Two evaluation studies. *Human Factors, 41,* 326–343.

Thornton, R. C., Kaempf, G. L., Zeller, J. L., & McAnulty, D. M. (1992). *An evaluation of crew coordination and performance during a simulated UH-60 helicopter mission* (Rep. No. ARI–RN–92–63). Fort Rucker, AL: U.S. Army Research Institute for the Behavioral and Social Sciences.

Thurman, R. A., & Dunlap, R. D. (1999). Assessing the effectiveness of simulator-based training. In *Proceedings of the 1999 Interservice/Industry Training Simulation and Education Conference.* Arlington, VA: National Training Systems Association.

Winkler, R. L., & Hays, W. L. (1975). *Statistics: probability, inference, and decision.* New York: Holt, Rinehart & Winston.