
CRM Procedures and Crew Performance Executive Summary of Advance Crew Resource Management (ACRM) *

Dr. Deborah Boehm-Davis, Dr. Robert Holt, Dr. Philip Ikomi, Jeff Hansberger, Jeff Beaubien, Kara Incalcaterra (George Mason University), Dr. Thomas Seamster (Cognitive & Human Factors), Captain William Hamman (United Airlines), and Captain Kim Schultz (Atlantic Coast Airlines)

December 1, 1998



Federal Aviation Administration
Office of the Chief Scientific and
Technical Advisor for Human
Factors, AAR-100

* This research was supported by the Office of the Chief Scientific and Technical Advisor for Human Factors (AAR-100) of the Federal Aviation Administration through FAA grant 94-G-034 to George Mason University. The opinions and conclusions expressed in this document represent solely the authors' viewpoints and do not represent official opinions of the FAA, the U.S. government, GMU, or any of the airlines participating in this research.

Introduction

Traditional Crew Resource Management (CRM) focuses on principles of coordination and communication, assuming that better crew communication and coordination improves crew performance and reduces errors. This has resulted in crew requirements that have been trained and assessed as additions to, rather than as part of, Standard Operating Procedure (SOP). The general approach to CRM is difficult to integrate into pilot training and certification. This study was undertaken to determine the effects of a more structured form of CRM based on CRM procedures.

The study began by identifying two main difficulties with a traditional CRM program at a regional airline. Operationally, crews had problems applying general CRM principles in specific operational situations. In assessment, evaluators could not provide crews with precise assessments of CRM performance because the principles were not clearly specified. To address these difficulties, CRM procedures were developed for normal as well as non-normal flight conditions.

There were several possible advantages to using CRM procedures. If the procedures are clearly specified, they could be more easily trained and applied than the general CRM principles. Further, the evaluation criteria for procedure performance would be more explicit and less subjective. Finally, a procedural approach would raise key aspects of CRM to the level of SOP, increasing the operational significance of CRM and providing crews with a more standard form of CRM. This study evaluated the extent to which a proceduralized version of CRM, called Advanced CRM (ACRM), would improve crew performance.

This evaluation dealt with the complexities of the operational environment by using a quasi-experimental design to compare a fleet with traditional CRM and ACRM training against a control fleet with traditional CRM training only. The pilots were assigned to fleets by normal airline rules and seniority with some pilots moving from one fleet to another during the evaluation. This resulted in some pilots in the traditional CRM fleet having experience with the specific ACRM procedures even though that fleet had not implemented ACRM. The main measures were crew performance in each fleet on yearly recurrent Line Operational Evaluations (LOEs) and on random Line Checks.

This paper presents a brief overview of developing ACRM (for a detailed presentation see Seamster, Boehm-Davis, Holt, and Schultz, 1998). Next, the assessment process is explained followed by the results of two performance evaluations. The first of those evaluations took place prior to the implementation of ACRM in order to establish baseline crew performance. The second evaluation was conducted after an intervening year of formal implementation and practice with ACRM procedures. In the second evaluation, the LOEs and Line Checks were

augmented with jump seat observations, an instructor survey, and a pilot survey to determine the full range of effects that ACRM had on pilot and crew performance.

Developing ACRM

ACRM is made up of a set of CRM procedures along with the process for developing, training, and assessing the effects of the procedures on crew performance. The ACRM process can be applied to any airline in order to proceduralize their CRM by embedding procedures in a range of crew activities in some or all phases of flight. ACRM procedures, as the one shown in this example, may be inserted into required crew briefings prior to critical times of flight such as takeoff and approach/landing. The following Takeoff Brief was designed to help crews address the main conditions relevant to each takeoff. The specific conditions that should be considered were listed under the instructions to Select and Prioritize.

Takeoff Brief
Statement of Condition
Select and Prioritize:
◇ Runway conditions
◇ Low visibility procedures
◇ Hydroplaning
◇ Crosswinds/windshear
◇ Terrain/MSA
◇ Aircraft performance
◇ Convective activity
◇ GPWS/TCAS alerts
◇ Fuel status/delays
Bottom Lines for takeoff
Backup Plan for takeoff
Initial heading and altitude

The ACRM process is used to develop and apply CRM procedures for an airline's unique operational context. The specific ACRM procedures evaluated in this study were developed and implemented at one regional airline.

The development of the ACRM procedures began with a review of an existing management system approach to CRM (Mudge, 1993), and a determination that some of those procedures did not fit the specific operational needs of the airline. Therefore, a tailored set of ACRM procedures was developed with the help of the airline to meet their operational needs. The new procedures had to be first approved by fleet managers and safety representatives. Subsequently, the fleet's normal checklist, Quick Reference Handbook (QRH), and Flight Operations Manual (FOM) were updated to incorporate those procedures and approved for use by the FAA official for the airline.

A first step in the tailoring process was to identify the airline's primary operational and safety concerns. Key issues were identified using the results from the National Transportation Safety

Board (NTSB) Safety Study, data from Aviation Safety Reporting System (ASRS) reports, and information from an instructor questionnaire. The NTSB Safety Study (NTSB, 1994) suggested a pattern to problems in the cockpit. Of the accidents analyzed in that report, 81% took place when the captain was the pilot flying (PF) and 73% occurred on the crew's first day of flight together. An analysis of ASRS incidents also suggested a pattern of crew problems associated with airlines flying the types of aircraft and missions common to regional operations. For example, the ASRS data suggested that crew distractions and problems in communication played a role in a large number of incident reports. Specific incident reports in areas such as "distractions" were reviewed to more precisely define the problems.

Information about specific airline problem areas was collected from a written survey asking instructors about crew observations they made during the previous year. This survey led to the identification of three problem areas for this airline: assertiveness, briefings, and decision making. In the area of assertiveness, the instructors felt that crewmembers were not speaking up with appropriate persistence. Operational briefings were not done consistently: 1) they were often too general, 2) they did not address specific conditions, and 3) there were no guidelines for what constituted an appropriate briefing. Finally, although crews discussed options when making decisions, they frequently did not develop clearly stated plans of action.

The data from these sources led to the development of three goals for developing CRM procedures:

- Reduce distractions to the pilot flying (PF) in both normal and abnormal situations so as not to interrupt cross checking and monitoring of the aircraft status.
- Provide structure to briefings with checklist format in order to enhance the crew's performance on the first day together and to improve transfer of critical information.
- Design checklists, the Quick Reference Handbook (QRH), and briefings to reduce workload and enhance decision-making skills, especially when crews would be fatigued, running late, or under high workload.

A design team consisting of the airline and research personnel used these goals to develop a set of specific ACRM procedures tailored for the involved airline.

Reliable Assessment of CRM

The Line Operational Evaluation (LOE) was used as the primary means of assessing CRM performance. LOE is an evaluation of individual and crew performance in a flight training device or flight simulator conducted during a real-time LOS under an approved AQP program as described in SFAR 58. The primary unit of both LOE design and CRM assessment is the *event set*, a group of related events that comprise the scenario and are inserted into a LOE session for specific training/evaluation purposes. The event set is a refinement of the Advanced

Qualification Program (AQP) concept of event, and is an integral part of training and evaluation. The event set is made up of one or more events, including an event trigger, a distracter, and supporting events. The event trigger is the condition or conditions under which the event is fully activated. The supporting events are other events taking place within the event set designed to further CRM and technical training objectives and add to overall realism. Finally, the distracter is a condition inserted within the event set time frame that are designed to divert the crew's attention from other events that are occurring or are about to occur.

To collect accurate crew performance data, a new LOE was developed, new assessment forms were implemented, and instructors were trained in how to administer the LOE and assess crew performance. This assessment process was developed to establish the rater reliability and validity required for a careful measurement of crew performance. This section presents the steps taken and the resulting rater reliability. The baseline crew performance and fleet performance differences due to ACRM are presented in the next two sections.

The new LOE presented crews with problems that taxed specific aspects of CRM and technical performance during a normal flight. To develop the new LOE, the airline first specified the technical and CRM objectives for the evaluation. Next, the airline and the research team jointly designed the LOE using the development process outlined in ATA (1994) ensuring that the content was realistic for their flight operations, conditions, and aircraft. Team members with aviation and research expertise facilitated the development of the LOE script, the construction of specific items to measure CRM and technical skills, and the development of the LOE worksheet. Assessments were based on event sets, made up of an event trigger, supporting conditions, and distracters. Event sets were used for this assessment in part because they have been shown to produce more reliable crew ratings compared with traditional overall session assessments (Seamster, Edens, & Holt, 1995).

Crew performance during the LOEs was assessed using structured worksheets. A corresponding LOE worksheet page guided the evaluation of each LOE event set. This page included success criteria that instructors could use to make more reliable ratings. After the LOE, the worksheet also helped instructors deliver a more balanced debrief that covered the CRM as well as the technical elements of each event set.

The specification of observable behaviors was an important element of the LOE worksheet. Each behavior was carefully identified and validated by instructors as being central to successful performance of the appropriate event set. These behaviors provided a point of focus for instructors during the observation of the LOE and subsequent debriefing. These focused observations were a basis for evaluating higher-level components of CRM and technical performance. Each evaluator synthesized the observed behaviors when rating the more general performance components on a standard rating scale.

The use of a standard rating scale across the full range of evaluation environments from LOE simulator sessions to Line Checks provided evaluators with extensive experience with the rating scale. Each point of the scale was based on a set of specified standards with both technical and CRM performance. The four-point scale covered the full range of possible crew performance from unsatisfactory to above standard. The labels and precise meanings of each point were defined as the result of discussion between instructors and researchers, and reinforced by appropriate training. For the LOE, this scale was reinforced by the task specific success criteria listed for each item.

The instructor training included information on the observable behaviors assigned in the LOE, the forms that were used to identify the specific event sets, the assessment criteria for the event sets, and video examples of volunteer line crews flying the LOE in a normal manner (not scripted). This practice enhanced the identification of focal behaviors and began the calibration process for the instructors. For optimal calibration, it was critical that instructors view video segments of event sets and discuss their ratings of the crew. After the instructors watched a video segment of the LOE and made their assessments, their ratings were processed by a data collection program and the results immediately returned so that they could compare their performance against the performance of the other instructors. This technique provided real numbers for comparison, removed some of the subjectivity from the instructor calibration process, and allowed refinement of the LOE guide, worksheet and success criteria.

In addition to the specific training for the LOE assessment, the instructor training also included the skills required to brief, administer, and debrief the LOE. Each instructor was required to 1) fly the LOE 2) observe the administration of the LOE by a previously trained instructor, and 3) practice giving the LOE as an evaluator. The instructors switched between pilot flying and instructor roles during the session to allow them the opportunity to experience the various roles in this type of training. As part of the training, instructors performed a facilitated debrief after each leg of the LOE. The instructors flying the simulator were encouraged to begin these debriefings with observations of the CRM skills that were exhibited or that should have been used during each event set. This training was designed to enhance observation and identification skills, give practice using the LOE worksheets for evaluation, and give practice in the type of crew-centered briefing/debriefing they would be conducting with normal line crews. The training was also designed to enhance each instructor's debriefing techniques by the sharing of concepts and techniques among the three pilots.

Baseline Fleet Performance

Once reliable and valid assessment was established, the next step was to determine how pilots in the two fleets were performing prior to implementing ACRM. Baseline averages for

ratings of crew performance were compared over a full range of LOE event sets. Equivalent performance between fleets was expected, but 18% of the items were significantly different across the two fleets. For these items, the control fleet with traditional CRM had higher averages than the fleet that was scheduled to have the additional ACRM training. This number of rated differences was more than would be expected by chance and suggested some source of differences between the fleets. These differences could have reflected a real distinction in performance between fleets, or they could have been due to instructors using less demanding standards for pilots in the traditional CRM fleet. Although this pattern of results could be attributed to either cause, subsequent analyses in the final year of evaluation supported the conclusion that different standards were used in the two fleets, which is discussed in the next section on performance differences.

One of the important lessons learned during this baseline assessment was that writing good items for observable behaviors was a difficult task. Similar to writing good attitude-measurement items, the observable behaviors had to be stated simply and unambiguously (Seamster, Boehm-Davis, Holt & Schultz, 1998). This experience helped the team write better items for observable behaviors and CRM skill evaluations when assessing performance differences due to ACRM implementation.

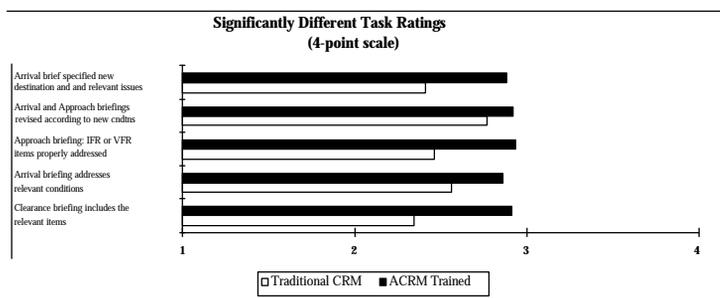
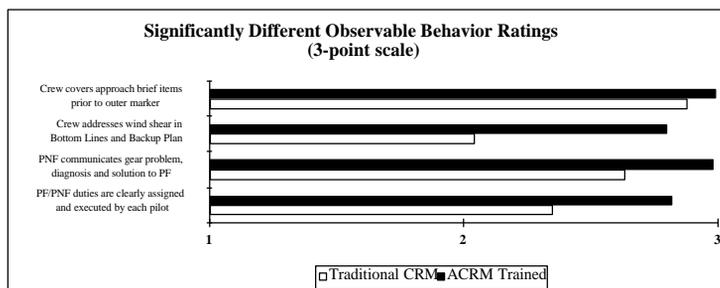
New procedures for inter-rater reliability training (IRR) were initiated at the airline during the first year. The development started with a simple spreadsheet implementation of agreement indexes that allowed immediate feedback and discussion among the instructors. As more problems in rater calibration were considered, IRR training gradually extended to include more diverse types of statistical analysis and feedback (see Williams, Holt, & Boehm-Davis, 1997). Instructor ratings of video taped performance were analyzed to determine their reliability and validity. Reliabilities for the observable behaviors, technical and CRM ratings, and PIC, SIC, and Crew ratings were in the .85 to .96 range. All reliabilities exceeded .70 which has been suggested as a lower limit for research use (Nunnally, 1967). The structural validity of the rating process (Holt, Johnson, and Goldsmith, 1997) was confirmed in that for most event sets the observable behaviors predicted a significant amount of variance in the technical and CRM ratings which in turn predicted a significant amount of the PIC, SIC, and Crew ratings.

Performance Differences Attributed to ACRM

To check on the possibility of different assessment standards in the two different fleets, four CRM evaluation items were evaluated twice by instructors from the traditional CRM fleet. In one evaluation, instructors were required to use the traditional CRM criteria for that fleet, and in the second evaluation they were required to use the ACRM criteria from the ACRM trained fleet. The results showed that when instructors used the ACRM criteria their ratings were significantly

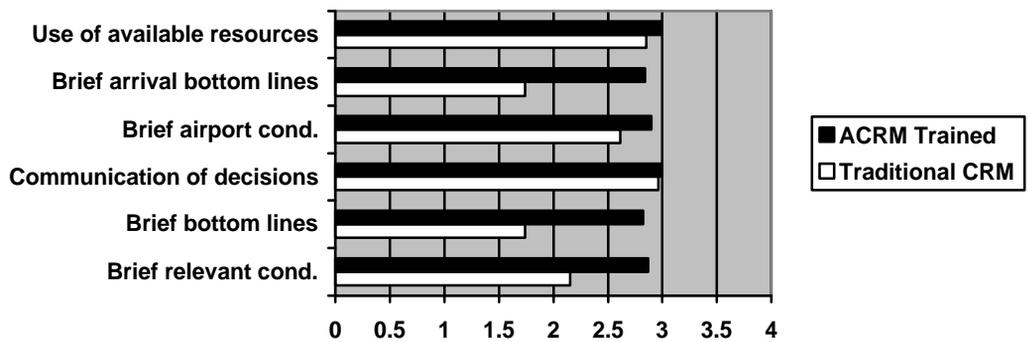
lower than when using the traditional CRM criteria. This demonstrated that there were different rating criteria for the two fleets. The difference between using traditional CRM grading standards and grading standards based on ACRM as SOP, was about half a scale point. This significant difference explains the occurrence of higher performance ratings for the traditional CRM fleet on some LOE items during the baseline year.

Since different fleet assessment standards were found, the crew performance analysis in the final year concentrated on items that could and would be assessed using a common standard across fleets. The LOE had ten items designed to be graded with the same standard in both the ACRM and traditional CRM-trained fleets. Four of those ten items were specific observable behaviors and the other six were more general task items. Nine out of ten items showed significantly higher scores for the ACRM fleet. This provided evidence that the ACRM fleet performed these observable behaviors and tasks consistently better and suggested that ACRM was having both a specific (observable behavior) as well as a general (task items) effect on crew performance.



To further examine the generality of the effects of ACRM training, the four observable behaviors and the five tasks were correlated with the other behaviors and tasks in the LOE for the ACRM-trained fleet. These nine items were significantly correlated with other items in the LOE to a degree greater than would be expected by chance. These results support a positive generalization of ACRM effects to other observable behaviors and tasks.

The revised Line Check form contained twelve items that could be affected by ACRM training and SOP implementation in the ACRM-trained fleet. Since both fleets used the same Line Check form, the mean performance on these items could be statistically compared. Six of the twelve items were significantly different, with the ACRM-trained fleet performing better than the traditional CRM fleet on all six. The combination of ACRM training and implementation into fleet SOP resulted a statistically significant performance difference between the fleets. For “communication of decisions” and “use of available resources” items, the observed differences are quite small. For the other four items, however, the difference in performance between the fleets is much larger. Clearly, the ACRM training plus SOP implementation can cause noticeable and important differences in crew behavior on Line Checks.



When comparable items with the same grading standards were examined across fleets, the LOE results clearly supported the effects of ACRM. The mean performance of the ACRM trained fleet was consistently and significantly better than the traditional CRM fleet. This result was also reflected in the Line Check data. The improved performance in both LOE and line operations suggests that the positive effects of ACRM did transfer from the training to the operational environment.

Supporting Evidence

To develop a complete understanding of crew performance, three additional measures were administered during the final year evaluation of ACRM. First, a jump seat observation was designed as an independent performance measure using a separate group of evaluators who assessed pilots from both fleets. Jump seat observations provided a converging measurement check on the performance differences of the two fleets using different sets of evaluators and new evaluation forms. Second, an instructor survey was developed for those who had trained pilots from both fleets. The survey asked instructors to compare the performance of ACRM-trained with that of traditional CRM-trained pilots during transition training. Third, a survey was developed for pilots to measure attitudes toward CRM and ACRM, knowledge and practice of

ACRM procedures, and perceived effects of ACRM. Attitude measurement was geared toward CRM in general and the ACRM procedures in particular. Knowledge included the concepts and details related to the new ACRM procedures. Practice was measured by how often pilots reported performing the new ACRM procedures and briefings in normal operations. Perceived effects measured the pilots' view of a variety of possible effects of the ACRM procedures.

Jump seat observations used the same performance criteria for rating crews from each of the two fleets. Crews were rated on their performance during each of the major phases of flight (take-off, cruise, and approach/descent) and on their overall flight performance. ACRM-trained pilots were compared with traditional CRM-trained pilots on their flight performance using a two-sample design that utilized individual item analyses to detect fine-grained differences between the two fleets. Independent sample t-test revealed that thirteen of the twenty items were significantly different, all in favor of superior flight performance in the ACRM fleet. Included in these thirteen items were the overall effectiveness items from the take-off, cruise, and approach/descent phases of flight and the overall effectiveness item referring to all phases of flight.

Even with a small set of 45 jump seat observations, the effects of ACRM were apparent. Due to the small sample, only the items on which observers perceived large differences between the fleets were significantly different. The average fleet difference across the thirteen significant items was .70, which is a rather large difference on a five-point scale. This large difference implies a practical as well as statistical difference in performance between the fleets. The ACRM-trained and ACRM-SOP fleet is noticeably superior in certain items reflecting departure, cruise, and arrival phases of flight.

The instructor survey asked for a comparison between pilots with ACRM training and those with traditional CRM training on the basis of the frequency and quality of specific behaviors. There were a total of 19 completed instructor surveys, and a factor analysis of their ratings resulted in three distinct factors; workload management, planning, and communication. For workload management, the combined ratings indicated that ACRM trained pilots were managing workload more frequently and with better quality than traditional CRM trained pilots. For planning, the ratings also showed that ACRM pilots plan more frequently and with better quality. For communication, the results similarly indicated that ACRM pilots communicated more frequently and with better quality. On these three basic factors, their evaluations of ACRM pilots versus pilots with traditional CRM were significantly in favor of ACRM pilots.

The pilot survey included knowledge of ACRM, practice of ACRM procedures, and attitudes toward CRM and ACRM. For pilots with just traditional CRM, the pilot survey only measured general attitudes toward CRM and specific attitudes toward ACRM. Those pilots were not asked

about detailed ACRM knowledge, practice of ACRM procedures, or the perceived effects of ACRM as those items could not be sensibly asked without the ACRM training and experience.

The pilot survey results also strongly support the ACRM program. ACRM-trained pilots had very positive attitudes toward CRM in general and ACRM in particular. When compared to a baseline, ACRM-trained pilots also showed significant knowledge of ACRM, frequently performed ACRM procedures, and overwhelmingly endorsed ACRM when it was put to the vote. The frequent performance of ACRM procedures was reassuring, and confirms the effectiveness of implementing these procedures as SOP for the fleet. Also, the endorsement votes for ACRM were overwhelmingly favorable, with over 90% responding positively to both items.

Convergence of Results. The possible effects of ACRM were examined with different evaluation methods, different samples of evaluators, and different samples of evaluated behaviors. The performance difference evidence and other types of supporting evidence *confirm* ACRM effects. The convergence of these different methods on showing positive effects of ACRM training and SOP implementation is compelling evidence that some ACRM effects did, in fact, occur.

Conclusions

The combination of appropriate ACRM procedures, ACRM training, and incorporation of ACRM into fleet SOP was effective in producing specific changes in crew performance. Different lines of evidence support the conclusion that the effects of ACRM training and SOP go beyond the specific procedures to improve overall crew performance. First, the items in the LOE for which the training makes a difference are positively correlated with a variety of other performance items across the event sets at a level greater than expected by chance. This supports the contention that ACRM effects are not just limited to specific items or just one event set. Second, the Line Check items also support a pattern of significant fleet differences in favor of the trained fleet. This demonstrates that the training and SOP implementation affected crew behavior on the line as well as in the simulator.

The jump seat observations offer the third line of evidence that the ACRM trained fleet showed better performance on items in all phases of flight. This supports the contention that ACRM effects are not just limited to a specific phase of flight. Fourth, the instructors perceived that pilots undergoing upgrade training or fleet transitions were performing significantly better when they came from the ACRM fleet than when they came from the traditional CRM fleet. The instructors judged that the pilots with ACRM background were better at workload management, effective cockpit communication, and planning. This supports the contention that the effects of ACRM are not just limited to one facet of pilot performance.

Fifth, the ACRM-trained pilots themselves perceived a wide variety of positive effects from this training. The twelve “expected effects” items covered a variety of possible effects of training. The results that all twelve of the items were perceived by the pilots to show positive effects of the ACRM training again supports a broad scope of effect for ACRM training and SOP implementation rather than a narrow scope. From the convergence of all these results, it seems that the effects of ACRM training and SOP implementation generalize to other aspects of crew performance.

It is more difficult to identify the relative contribution of the ACRM training course versus the impact of the new procedures. To determine the relative contribution of these two factors would require a fleet in which ACRM would be trained in a classroom without it being implemented as SOP, and a fleet in which ACRM would be implemented as SOP without a separate training course. Since pilots who only had the training showed the most positive attitudes to ACRM, just the training component may be one source of attitude change. On the other hand, implementing the procedures as SOP implementation may lead to a more uniform and consistent pattern of behavior across the pilot population as well as across the different phases of flight.

There are several practical implications of these results. The main implication is that turning specific aspects of CRM into procedures can improve crew performance both in the simulator and on the line. Appropriately designed and trained procedures can enhance the crew’s ability to communicate effectively, plan, manage workload, and solve problems during flight operations. This does not, however, mean that CRM procedures can resolve all crew performance problems. Procedures must be carefully designed and implemented to fit the operational context, the types of pilots employed by the airline, and the corporate climate (Seamster, Boehm-Davis, Holt, & Schultz, 1998). Alternatives to CRM procedures, such as training interventions or technical fixes, must also be considered. Further, the appropriate design and implementation of procedures.

Where appropriate, developing and training CRM procedures may have significant advantages over training general CRM principles. Training CRM principles puts the burden on the pilot to figure out how and when to implement the principles in the cockpit. CRM procedures, when properly developed under a process such as ACRM, remove this burden by specifying the time and manner of implementing the CRM principle. Increasing the ease of implementation should improve the frequency of CRM performance. Similarly, improving the standardization of the interaction should improve the effectiveness of crew performance, particularly in communication and coordination among the crew. Further, idiosyncratic differences in the implementation of CRM principles across crews may be reduced by the standardization afforded by CRM procedures. For example, some Captains under traditional

CRM may consider a thorough briefing important for take off but not landing, whereas other Captains may have the opposite view. Such potential differences in briefing behavior may be reduced when the CRM principle is made into a procedure enacted at specific phases of flight. Finally, requiring CRM procedures across phases of flight might be one way to ensure standard and predictable CRM behavior throughout the flight.

References

- ATA (1994). *Line Operational Simulations: LOFT Scenario Design, Conduct and Validation*. LOFT Design Focus Group, AQP Subcommittee Report, Air Transport Association, November 2, 1994.
- Holt, R.W., Johnson, P.J., & Goldsmith, T.E. (1997). Application of psychometrics to the calibration of air carrier evaluators. In *Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting*, (pp. 916-920).
- Mudge, R.W. (1993). Pilot judgment - and the management system. *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 216-220). Columbus, OH: The Ohio State University.
- NTSB (1994). *Safety Study: A Review of Flightcrew-Involved, Major Accidents of U.S. Air Carriers, 1978 through 1990* (PB94-917001 NTSB.SS-94/01). Washington DC: National Transportation Safety Board.
- Nunnally, J.C. (1967). *Psychometric Theory*. New York: McGraw-Hill.
- Seamster, T.L., Boehm-Davis, D.A., Holt, R.W., & Schultz, K. (1998). *Developing Advanced Crew Resource Management (ACRM) Training: A Training Manual*. Washington, DC: Federal Aviation Administration, AAR-100.
- Seamster, T.L., Edens, E.S., & Holt, R.W. (1995). Scenario event sets and the reliability of CRM assessment. *Proceedings of the Eighth International Symposium on Aviation Psychology*, (pp. 613-618). Columbus, OH: The Ohio State University.
- Williams, D.M., Holt, R.W., & Boehm-Davis, D.A. (1997). Training for inter-rater reliability: Baseline and benchmarks. *Proceedings of the Ninth International Symposium on Aviation Psychology* (pp. 514-519). Columbus, OH: Ohio State University.