An Evaluation of the Rating Process used by Instructor/Evaluators in a Line-Operational Simulation: Preliminary Evidence of Internal Structure Validity

J. Matthew Beaubien Robert W. Holt

George Mason University

Captain William R. Hamman

United Air Lines

January 21, 1999

Technical Report #98-002

Please address all requests for reprints to:

Dr. Robert W. Holt
ARCH Lab
MSN 2E5
Department of Psychology
George Mason University
Fairfax, VA 22030-4444
(703) 993-1344
bholt@vms1.gmu.edu

Abstract

Crew Resource Management (CRM) training programs have existed for more than a decade, yet relatively few attempts have been made to assess their effectiveness using methodologically rigorous designs. Of the studies that do exist, most are summative in nature. Curiously, much less research has examined the specific processes used by instructor/evaluators (I/Es) when making their evaluations of crew-level CRM and technical proficiency. In the current study, data were collected from two separate Line Oriented Evaluations (LOEs) in order to compare instructor/evaluators' rating processes with the carrier's standard operating procedure (SOP). The data suggest that instructor/evaluators were using the rating process as designed. Furthermore, the data also suggest that it is indeed possible to link crew-level evaluations of CRM proficiency with specific, behavioral indicators. Implications and directions for future research are discussed.

Authors' Notes

This technical report has been written with the needs of both academics and applied researchers in mind. However, because the material is of a complex nature, it will be difficult to satisfy the informational needs of both audiences. For example, academics may feel that the material is not presented in enough detail, while applied researchers may feel quite the opposite. Therefore, to make future reports more "user-friendly", the authors welcome any comments that you might have. Please direct your comments to the address listed on the front cover.

The authors would also like to acknowledge the support of several individuals who assisted in the generation of this manuscript. Without their help, this research effort would not have been possible. Specifically, the authors extend their sincerest thanks to Dr. Thomas Longridge (FAA, AFS-230), Dr. Eleana Edens (FAA, AAR-100), the Quality Assurance department at United Air Lines, Dr. Thomas Seamster (Cognitive and Human Factors), and Captain Kim Schultz (Atlantic Coast Airlines).

Funding for this report has been provided by a grant from the Federal Aviation Administration entitled “Analysis of CRM Procedures in a Regional Air Carrier”. However, the views expressed in this report are those of the authors, and do not necessarily reflect the opinions of their sponsoring agencies. Neither the authors nor their sponsoring agencies make any warranty, expressed or implied, or assume any legal liability for any third party's use, or the results of such use, of any information, apparatus, product, or process discussed in this report.

Key Terms and Acronyms

Bartlett’s Test of Sphericity - A statistical test that determines whether or not there is a sufficient number of non-zero correlations in a correlation matrix to warrant the use of exploratory factor analysis. According to Tabachnick & Fidell (1996), the test may exhibit significant results with large samples even if the item inter-correlations are rather low.

Crew Resource Management (CRM) - The effective use of all resources (human, informational, and hardware) on the flight deck.

CRM Performance - An overall, crew-level rating of leadership behaviors, teamwork skills, situational awareness, interpersonal communication, and the utilization of all available forms of information during the simulated flight. Effective CRM performance is hypothesized to be correlated with, but not identical to, effective technical performance.

Event Set (ES) - A relatively independent segment of a simulated flight. Each event set typically includes a trigger, possible distractors, and environmental conditions. By segmenting the flight into more manageable units, event sets facilitate the evaluation of crew performance in the simulator.

Exploratory Factor Analysis (EFA) - A data reduction technique that reduces the correlations among a set of variables to a more parsimonious subset (for interpretation purposes). This technique is conceptually similar to principal components analysis. However, unlike principal components analysis, EFA analyzes only the shared variance among a set of items; unique variance for individual items is ignored. Typically, the general term "factor analysis" is used to describe both exploratory factor analysis and principal components analysis.

Intra-Class Correlation (ICC) - A statistical procedure that is used to assess the relative amount of between- and within-group variance in a measure. Intra-class correlations are scored on a metric which ranges from zero to one. A value of zero indicates that all the observed variance resides within groups (e.g., there is no between-groups variability). Conversely, a value of one indicates that all of the variance resides between groups (e.g., there is no within-groups variability). Typically, values greater than .80 suggest that a variable is measured at the "group" level, as only 20 percent of the observed variance can be attributed to individual differences within groups.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy - A statistical test that determines whether or not there is a sufficient number of non-zero correlations in a correlation matrix to warrant the use of exploratory factor analysis. Specifically, the measure of sampling adequacy compares the sum of the squared correlations to the sum of the squared correlations plus the sum of the squared partial correlations. As the partial correlations become smaller, the measure of sampling adequacy becomes closer to a value of one. According to Tabachnick & Fidell (1996), values greater than .60 are generally considered acceptable.

Line Oriented Evaluation (LOE) - An evaluation of individual and crew performance in a real-time, full-motion flight simulator, during which a flight is simulated from take-off to landing. In an LOE, emphasis is placed on performance evaluation, rather than training. Although both CRM and technical flight skills are evaluated during an LOE, there is a somewhat stronger emphasis on CRM proficiency.

Line Operational Simulation (LOS) - The generic term for real-time, full-motion simulated flights. Simulated flights of this nature can be used for performance evaluation (Maneuver Validation, Line Oriented Evaluation), training (Line Oriented Flight Training), or other purposes (Special Purpose Operational Training).

Meta-Analysis - An empirical cumulation of the existing data regarding a single bivariate relationship. Typically, the observed correlation from each study is corrected for range restriction and measurement unreliability. The correlations are then weighted by the study's sample size. Finally, the sample-weighted correlations are averaged across studies. The resulting value is often considered a reasonable estimate of the true bivariate relationship in the total population.

Observable Behaviors - A set of specific tasks (on an LOE evaluation form) that a crew is expected to perform if they are to meet the challenges posed in a given event set. While the list of observable behaviors is meant to represent the most important/common behaviors for a given phase of flight, it is by no means exhaustive.

Path Analysis - A series of multiple regression analyses that are used to test the plausibility of a causal model. At each stage of the analysis, the dependent variable is regressed on all predictors that are hypothesized to exert direct effects. Typically, path analyses are used to test whether the effect of an independent variable on a dependent variable is fully- or partially-mediated.

Pilot in Command (PIC) - An overall, individual rating of the captain's performance in the simulation, regardless of whether the captain is physically controlling the aircraft. PIC ratings are hypothesized to be largely a function of the crew-level CRM ratings, as the captain is primarily responsible for providing leadership behaviors, setting the tone of the cockpit, and initiating crew briefings. PIC ratings are also hypothesized to be a function of technical proficiency, although to a lesser extent.

Second in Command (SIC) - An overall, individual rating of the first officer's performance in the simulation, regardless of whether the first officer is physically controlling the aircraft. SIC ratings are hypothesized to be more influenced by CRM rather than technical proficiency, although to a lesser extent than the PIC ratings (as the first officer typically takes a less active role in managing the cockpit).

Principal Components Analysis (PCA) - A data reduction technique that is conceptually similar to exploratory factor analysis. Unlike exploratory factor analysis, however, PCA analyzes <u>all</u> the variance in a correlation matrix. No attempt is made to partition variance into "common" and "unique" components. According to Tabachnick & Fidell (1996), if the variables in a correlation matrix contain little unique variance, the results from EFA and PCA will be similar.

Technical Performance - An overall, crew-level evaluation of behaviors that are directly related to the physical operation of the aircraft. Technical performance is often referred to as "stick-and-rudder" proficiency. An example would be the ability to land an aircraft under conditions of high wind shear. Technical performance is hypothesized to be a necessary but insufficient precursor to effective CRM performance.

Topic-Level Ratings - Topic-level ratings are conceptually similar to observable behaviors, however they are specifically written to be less specific. Because topic-level ratings can be interpreted more broadly by the instructor/evaluators (compared to observable behaviors), they are presumed to exert fewer cognitive demands on the rater, thereby increasing the accuracy of the rating process. Topic-level ratings were introduced as an alternative to observable behaviors.

Varimax Rotation - A statistical technique that is used to aid in the interpretation of a factor analysis/principal components analysis solution. Varimax procedures rotate the factor axes such that all variables tend to load high on one factor, and low on all of the others. Before using a varimax rotation technique, it is incumbent on the researcher to show, either by means of theory or empirical data, that the underlying factors are relatively independent of one another.

Evaluation of the Rating Process used by Instructor/Evaluators in a Line-Operational Simulation: Preliminary Evidence of Internal Structure Validity

In the United States, commercial aviation remains the safest form of mass transportation. For any given flight, it is estimated that the probability of survival is approximately 99.99 percent (National Transportation Safety Board, 1994). Nevertheless, when accidents do occur, the results are often disastrous. For example, at the time this document was being prepared, SwissAir flight 111 had recently crashed off the coast of Nova Scotia, killing all 229 on board.

Given the precision and reliability of modern jet aircraft technology, mechanical causes of aviation accidents are quite rare (see Helmreich & Foushee, 1993, for a review). Analyses of archival data suggest that the major cause of aviation accidents is human error on the flight deck (Boeing Commercial Aircraft Group, 1994; National Transportation Safety Board, 1994). To reduce the incidence of human error, and by extension the number of accidents, the Federal Aviation Administration (FAA) has recommended the implementation of Crew Resource Management (CRM) training programs (Federal Aviation Administration, 1993).

Crew Resource Management Training

CRM training programs have been designed according to the principles of Human Factors, a multi-disciplinary field that explores the interface between humans and machines in complex systems. The purpose of CRM training programs is to provide trainees with the knowledge and skill to effectively mange all available resources, whether they be human resources, hardware resources, or informational resources (Federal Aviation Administration, 1993). While early CRM programs focused exclusively on the behavioral styles of individual crew members (Lauber, 1984), recent advances in CRM training have expanded their scope to include crew interactions with air traffic control personnel, dispatchers, and maintenance technicians, as well as the proceduralization of CRM skills with briefings, checklists, and memory items (Helmreich et al., in press; Seamster et al., 1998). In general, CRM training programs target three main knowledge/skill clusters: communication processes and decisions; team building and maintenance; and workload management and situational awareness (Federal Aviation Administration, 1993; Gregorich & Wilhelm, 1993). These skills are typically trained using a combination of methods, such as lecture, group discussion, and role play.

Evaluation of CRM Training Interventions

Unfortunately, relatively few studies have assessed the effectiveness of CRM training interventions using methodologically rigorous designs. This is due, in part, to a number of operational and statistical constraints associated with conducting large-scale field studies in the aviation domain. In general, validation studies of CRM training typically employ four types of criterion measures: archival reports of aviation accidents/incidents, "objective" accident data, self-reported crew attitudes, and ratings of crew performance in Line-Operational Simulation (LOS) environments. Unfortunately, each data source is associated with its own unique problems.

Archival Reports. While NASA, the FAA, and individual carriers independently maintain archival databases that contain narrative descriptions of aviation accidents/incidents, such data is typically not amenable to statistical analysis. This is due to the fact that each accident/incident occurs in a unique, multi-factor situation. Unfortunately, however, narrative reports are typically non-standardized, thereby precluding meaningful comparisons among accidents/incidents (Kanki &

Palmer, 1993). Furthermore, narrative reports are typically de-identified for security reasons (Helmreich & Foushee, 1993). These constraints make it virtually impossible for researchers to probe for follow-up data, such as by linking accident/incident information with CRM training performance, measures of organizational climate, or other relevant factors. Finally, because such reports are made on a voluntary basis, they may reflect a biased sub-sample of those incidents which occur every year (Helmreich, Merritt, & Wilhelm, in press; Kanki & Palmer, 1993).

Objective Accident Data. There are also a number of operational problems associated with the use of "objective" accident data as indicators of CRM training effectiveness. Given the low base rate of aviation accidents, it would be necessary to collect data over the course of several years before a reasonable sample size could be accumulated (Helmreich & Foushee, 1993; National Transportation Safety Board, 1994). This is problematic given the fact that the industry is in a constant state of change, resulting in part from corporate mergers and the continual introduction of new technologies on the flight deck. As a result, statistical results could be either masked/confounded by these changing conditions, thereby rendering the findings inconclusive (Cook & Campbell, 1979).

Furthermore, given the time required to perform such a study, environmental conditions could have changed to such a degree as to make the particular CRM training intervention obsolete even before the evaluation had been completed. Ultimately, the evaluation of CRM training interventions must be performed under relatively aggressive time frames, in order to provide results that are both statistically interpretable and useful to carrier personnel.

Crew Attitudes. Given these constraints, much of the existing CRM research has focused on crew attitudes towards CRM principles and practices. For example, research conducted by Helmreich and colleagues (Helmreich, 1991; Helmreich & Foushee, 1993; Helmreich & Wilhelm, 1991) has consistently shown that crew members perceive CRM training as being both useful and relevant to the operation of the flight deck. Nevertheless, it must be noted that crew members' attitudes towards CRM training do not necessarily imply the effective implementation of CRM behaviors on the flight deck. While Helmreich and colleagues correctly note this limitation, they do suggest that crew attitudes are an essential first step in the evaluation of CRM training programs (Helmreich, 1984; Helmreich & Foushee, 1993).

Although this argument is intuitively appealing, recent research suggests that this optimism may be misplaced. For example, Alliger et al.'s (1997) meta-analysis of training effectiveness suggests that measures of perceived utility correlate only .26 with measures of immediate, declarative knowledge, and .03 with changes in workplace behaviors. Therefore, there is substantial evidence to suggest that attitudes toward CRM training may not be related to behavioral transfer on the fligh deck. Furthermore, questions remain regarding the long-term stability of crew attitudes towards CRM principles (Helmreich & Wilhelm, 1991). Given the limitations of attitude and perceived utility measures as predictors of line performance, many researchers have begun to explore more behaviorally-based estimates of CRM training effectiveness.

Ratings of Crew Performance. To date, however, only two major studies have empirically assessed the effectiveness of CRM training using crew performance data as the criterion of interest. The first was performed by Clothier (1991). Using a team of trained check airmen and academic researchers, Clothier demonstrated consistent, positive relationships between crew attitudes and performance ratings on fourteen separate measures of group process behavior.

While encouraging, Clothier's results leave a number of unanswered questions. First, given the nature of the evaluation form/technique, it is impossible to rule out alternative explanations, such as halo error and/or priming effects, for the observed differences in mean performance ratings (Cook & Campbell, 1979). Second, the author only presented the results of statistical significance tests, rather than their associated measures of effect size. Given the relatively large sample size employed, the statistical significance tests were powerful enough to detect extremely small differences in performance ratings (Cohen, 1988); differences which, although statistically significant, may not be practically different from zero. Finally, the author's results only addressed the notion of mean differences between the control and experimental groups. Much less emphasis was focused on the specific behavioral processes that crew members performed, and the relationship between these processes and overall crew-level ratings of performance. Yet, by ignoring the underlying processes, it is difficult to discern how crews use CRM principles on the line, or where future interventions should be targeted to improve performance.

The second major study was performed by Holt et al. (1998). This multi-sample study evaluated the effectiveness of a proceduralized CRM training intervention (ACRM) using converging operations (see Seamster et al., 1998, for a review of ACRM training). First, trained and untrained crews were compared using comparable items and rating standards in an LOE environment. Second, instructor/evaluators were surveyed regarding their overall impressions of the performance of trained and untrained crews, with particular emphasis on crews that had transitioned from one fleet (the fleet which did not receive training) to the other (the fleet which had received training). Third, jump-seat evaluations of typical performance measures were collected. Although each study, when considered individually, contained methodological flaws, taken together the three lines of research suggest that the ACRM-trained crews performed both statistically and practically better than their untrained counterparts.

The Current Study

Both the Clothier (1991) and Holt et al. (1998) studies employed between-group designs in which trained crews were compared to untrained crews. Nevertheless, it is often desirable to make within-fleet comparisons, for example when all crews have previously received CRM training. In such studies, the emphasis shifts from the evaluation of crew performance to an assessment of the rating process used by instructor/evaluators. For example, such analyses allow carrier personnel to compare the instructor/evaluators' actual rating processes to the carrier-specific standard operational procedure (SOP). At the same time, such analyses allow carrier personnel to identify the relative strengths and weaknesses of a given LOE, in order to improve the designs of future LOEs. To date, however, no published studies have addressed this issue.

Furthermore, the current study represents an initial attempt to address several gaps in the CRM evaluation research base. First, heeding Gregorich and Wilhelm's (1993) call for more research concerning the impact of CRM training on crew behavior, ratings of actual crew performance were conducted in a full-motion simulator. To reduce the impact of rater-induced errors (e.g., halo error, priming effects), accountable measures were used (e.g., the evaluator was required to justify his/her ratings to the crew) and crew process data were collected. Second, measures of effect size were used to offset the problems typically associated with traditional null-hypothesis significance tests. Finally, data were collected using two separate samples, in order to cross-validate the observed findings, and to make modifications to the original research design.

The analytical techniques presented in the current study were specifically chosen to serve as basic data quality checks that all carriers can perform to determine if their evaluation process is occurring as designed (e.g., carrier SOP). If the data suggest that the evaluation is proceeding as designed, then carrier personnel can have greater confidence when analyzing the data using other statistical techniques. However, if ratings are not being made in accordance with SOP, then carrier personnel should be wary when interpreting the results of statistical tests based upon these data.

Method

This report analyzes the results of two field studies that were conducted approximately one year apart. Both studies involved the assessment of crew performance in a Line Oriented Evaluation (LOE) setting. The Line Oriented Evaluations were designed according to standard regulatory and industry practices (Federal Aviation Administration, 1990; Prince et al., 1993). Specifically, each LOE was designed to simulate an actual flight from take-off to landing, during which crew members assumed their typical flight roles. Each simulated flight was decomposed into six separate event sets (ESs). Each event set represented a distinct phase of flight and included an environmental trigger, specific behaviors that the crew were expected to perform, and a set of pre-defined rating criteria. It must be noted that the event sets were developed for the evaluators' purposes only: to the crews performing in the simulator, the LOE operated as an uninterrupted flight.

During the LOE, the instructor/evaluators concurrently performed three separate roles. First, they interacted with the crew by role-playing the air traffic controller (ATC). Second, they manipulated the physical conditions of the simulator, such as weather and physical malfunctions of the aircraft. Finally, they evaluated the crew using a standardized evaluation form. Upon completion of the LOE, the instructor/evaluator facilitated a debriefing of the crew members' performance. These debriefings linked the instructor/evalator's performance evaluations with specific, behavioral examples by means of crew self-critique, instructor feedback, and the observation of videotaped performance examples. At the completion of the debrief session, videotapes of crew performance were erased according to standard industry and regulatory practices (Federal Aviation Administration, 1990).

Setting

Both studies were performed at the same international air carrier. All crews were line pilots in the Boeing 757 fleet, and were evaluated in a full-motion Boeing 757 aircraft simulator. This simulator is a realistic facsimile of a Boeing 757, and provides both real-time scrolling video as well as simulated movement. As stated earlier, the crew members performed their typical flight duties during the simulated flight. While this was occurring, the instructor/evaluators manipulated the simulator, interacted with the crew, and evaluated their performance.

As the LOE is a “jeopardy” evaluation, the crews' performance on the LOE determined their flight status. For example, an LOE failure would typically result in an individual/crew being removed from operational flight duties for one month. All participants were expected to exert maximal effort, which would classify the LOE as a measure of maximal, rather than typical, task-related performance (Dubois et al, 1993; Sackett et al., 1988). Therefore, we can expect the instructor/evaluators' ratings of crew performance to reflect the crew members' maximal ability levels, although it is recognized that the observed results may not necessarily generalize to typical performance on the line. Nevertheless, the primary emphasis of this study concerned an evaluation of the rating processes used by instructor/evaluators in an LOE environment. As a result, the generalization to crew performance on the line remains an important, although secondary issue.

Participants

For this carrier, the Boeing 757 fleet had been trained under the FAA's Continuing Qualification Program (CQP). Therefore, all crew members had received previous training in CRM principles. Furthermore, virtually all crew members had previously been evaluated in LOE environments. Unfortunately, because of operational issues, it was impossible to link Pilot Identification Numbers (PIN numbers) to organizationally-maintained databases that contain other crew-related information. Therefore, demographic characteristics of the crews, such as their mean number of flight hours, could not be calculated.

During the first LOE (herein referred to as LOE Alpha), 636 crews were evaluated. Approximately one year later, 837 crews were evaluated in the second LOE (herein referred to as LOE Bravo). Because all Boeing 757 crew members are required to undergo recurrent evaluation on a yearly basis, it is likely that there is substantial overlap between the two samples. Unfortunately, because pilot PIN numbers were generated randomly each year, it was impossible to link the two data sets. Therefore, the exact proportion of overlap between the two samples is impossible to determine. As a result, these two samples should not be considered completely independent. However, the two LOEs were dissimilar with regard to content, as they provided the crews with unique challenges that required substantially different responses. This should serve to reduce the impact of testing effects on the observed results (Cook & Campbell, 1979).

Materials

Data were collected by the instructor/evaluators using standardized rating forms. For each event set, the instructor evaluator made three sets of ratings. The first set of ratings consisted of either "observable behaviors" (LOE Alpha) or "topic-level ratings" (LOE Bravo). Both sets of ratings were made on a three point rating scale (1 = not observed, 2 = partially observed, 3 = fully observed), and were developed by subject matter experts to reflect the knowledges, skills, and abilities (KSAs) required to successfully face the challenges for that particular event set. Observable behaviors and topic-level ratings are conceptually similar to one another, with the exception that topic-level ratings are somewhat more general than observable behaviors. For a detailed comparison of topic-level ratings and observable behaviors (matched for event set content), see Table 1.

Insert Table 1 about here

The switch from an "observable behavior" rating system (LOE Alpha) to a "topic-level" rating system (LOE Bravo) was made based upon operational considerations. Specifically, during the first year of data collection, it was noted that the empirical relationships between observable behaviors (LOE Alpha) and the other performance ratings were smaller than anticipated. Anecdotal evidence, based on interviews with instructor/evaluators, suggested that the low validities were due to the high level of cognitive workload faced by the instructor/evaluators in the simulator. To address this problem, during the second year, fewer, more general "topic-level" ratings (LOE Bravo) were substituted for the observable behaviors on the LOE assessment form. Nevertheless, both observable behavior and topic-level ratings were assessed as crew level variables. These behaviorally-based ratings could be performed by either crew member, and regardless of which crew member(s) performed the observable behavior/topic, both crew members were assigned the same rating.

The second set of ratings consisted of crew-level ratings of CRM and technical performance. These ratings were made on a four-point scale (1 = repeat required, 2 = below standard, 3 = corporate standard, 4 = above standard). The processes for rating crew-level CRM and technical performance ratings were identical across both LOEs. According to the carrier’s standard operating procedure, crew level CRM and technical performance ratings were to be contingent upon the crew members' behaviorally-based (observable behaviors or topic-level, depending on the LOE under consideration) performance ratings. As before, both crew members were assigned the same CRM and technical performance ratings for a given event set.

The third and final set of ratings consisted of individual-level ratings of Pilot in Command (PIC) and Second in Command (SIC) performance. While these ratings were conceptualized at the individual level of analysis, anecdotal evidence suggests that they may contain a sizable crew-level component. This is due to the fact that successful crew performance requires the effective coordination of effort between the captain and the first officer. For example, it would not be unexpected to observe a poor first officer "downgrading" his/her captain's performance rating, because the captain was forced to compensate for the first officer's deficiencies. Therefore, the level of analysis for PIC and SIC ratings was empirically tested using intra-class (ICC) correlations (Howell, 1994).

For each event set, intra-class correlations between PIC and SIC ratings were computed. In general, ICC values typically exceeded .70, suggesting a substantial degree of "crewness". While this is undesirable from an organizational perspective (as these ratings were designed to assess individual proficiency), this has desirable statistical properties due to the fact that all variables are measured at the same level of analysis.

The Rating Process

Carrier standard operating procedure dictates that for each event set, instructor/evaluators are to use behavioral indices (observable behaviors or topic-level ratings) when making their crew-level ratings of CRM and technical performance. While these behavioral indices are not meant to be an exhaustive list of the required behaviors for each event set, they have been systematically developed by subject matter experts to reflect the most common/important areas of behavior for the particular phase of flight.

Next, the carrier standard operating procedure suggests that crew-level CRM and technical performance ratings are to be used in making overall ratings for the captain (PIC) and first officer (SIC). As all ratings can be conceptualized at the crew level, path analytic techniques (Cohen & Cohen, 1983; Pedhazur, 1982) were used to compare the rating processes used by instructor/evaluators to the corporate standard. Quite simply, if the instructor/evaluators were making their ratings according to the standard operating procedure, it would be expected that, at each stage of the path analysis, the predictor variables should account for both statistically and practically significant amounts of variance in the criterion variables. The standard operating procedure for rating crew performance in the simulator is presented diagrammatically in Figure 1.

Insert Figure 1 about here

As stated earlier, measures of effect size were estimated using standard path analytic techniques. First, crew-level CRM and technical performance ratings were separately regressed onto the behavioral indices (either the observable behaviors or topic-level ratings, depending on the LOE under consideration). Next, Pilot in Command (PIC) and Second in Command (SIC) ratings were separately regressed onto crew-level ratings of CRM and technical proficiency. At each stage, the overall amount of variance accounted for was assessed by R^2 measures of effect size. Individual paths were indexed by means of squared semi-partial correlations, and were tested for statistical significance using a conservative cutoff value ($\underline{p} < .01$). Finally, the models were tested using Baran and Kenny's (1986) three-step approach to determine if they represent partially or fully mediated relationships.

Given that each LOE contained six event sets, twelve separate path analyses were performed. Means, standard deviations, and correlation matrices for the six event sets (LOE Alpha) are provided in Tables 2 though 7.

Insert Tables 2 through 7 about here

Results

Before testing the congruence between the actual ratings and the carrier's standard operating procedure for the first year of data collection, preliminary data checks were made via examination of descriptive statistics and exploratory factor analyses. It should first be noted that the sample sizes were severely truncated due to the presence of missing data. For example, although over 600 crews were assessed, listwise deletion procedures reduced our sample by almost fifty percent.

As can be seen in Tables 2 through 7, all variables were negatively skewed. Based on statistical theory, this range restriction should attenuate the observed correlations and measures of variance explained (Howell, 1994). There is some question as to whether this range restriction represents a statistical artifact or a true organizational phenomenon. As the LOE ratings were made for administrative purposes, previous research would lead us to expect some form of leniency bias (Kozlowski, Chao, & Morrison, 1998; Murphy & Cleveland, 1995). On the other hand, given that all crew members were subject to rigorous selection techniques, had extensive experience/training, and were familiar with LOE evaluation procedures, it would also be expected that most crew members would meet or exceed the minimum carrier benchmark of proficiency. In either case, normally distributed performance data were not expected (Murphy & Cleveland, 1995). As it was impossible to disentangle these two competing hypotheses, corrections for range restriction were not performed.

Factor Structure. Next, principal components analyses were performed to better understand the underlying data structure. First, the thirty-two observable behaviors (across all six event sets) were factor analyzed using principal components analysis with a varimax rotation[1]. Preliminary data quality checks suggest that the correlation matrix contained a substantial number of non-zero correlations, and as such, was amenable to exploratory factor analyses. For example, the Kaiser-Meyer-Olkin measure of sampling adequacy yielded a value of .815, and Bartlett's test of sphericity yielded a value of 5816.242 (df = 496, $\underline{p}$ < .000). Both values fall within appropriate numerical ranges, as suggested by Tabachnick & Fidell (1996). Principal components analyses suggested the presence of ten components which accounted for approximately 62 percent of the observed variance. The first component, which accounted for 21 percent of the observed variance, loaded highly on observable behaviors from the "Descent" and "Approach" event sets. However, because of high cross-loadings, the remaining components were virtually uninterpretable.

Next, principal components analyses were performed on the twelve crew-level CRM and technical performance ratings (across the six event sets). Again, preliminary data quality checks suggest that the correlation matrix was amenable to factoring. The Kaiser-Meyer-Olkin measure of sampling adequacy yielded a value of .798 and Bartlett's test of sphericity yielded a value of 4769.678 (df = 66, $\underline{p}$ < .000). Again, these values fall within acceptable mathematical ranges. Three principal components emerged, accounting for approximately 57 percent of the observed variance. In general, all items loaded on one, and only one component; a single item had a cross-loading greater than .30. The first component included the CRM and technical ratings from the "Cruise" and "Descent" event sets. The second component included the CRM and technical ratings from the "Pre-Departure" and "Taxi-Out" event sets. The final component included the CRM and technical ratings for the "Climb" and "Approach" event sets.

Finally, principal components analyses were performed on the twelve PIC and SIC ratings (across the six event sets). Again, the correlation matrix exhibited desirable statistical properties, suggesting that it was amenable to exploratory factor analyses. The Kaiser-Meyer-Olkin measure of sampling adequacy yielded a value of .743 and Bartlett's test of sphericity yielded a value of 6447.838 (df = 66, $\underline{p} < .000$). The factor structure which emerged was quite similar to the one for the CRM and technical performance ratings. However, this time four principal components emerged. These four components account for approximately 68 percent of the observed variance. As in the previous analysis, the first component included the PIC and SIC ratings from the "Cruise" and "Descent" event sets. Again, the second component included the PIC and SIC ratings from the "Pre-Departure" and "Taxi-Out" event sets. The third component, however, included only the items "Approach" event set. The fourth and final component was comprised of PIC and SIC ratings on the "Taxi-Out" and "Climb" event sets.

Path Analyses. Finally, the instructor/evaluators' rating process was compared to the carrier's standard operating procedures using path analytic techniques. The path analyses were performed according to specifications set forth by Cohen & Cohen (1983) and Pedhazur (1982). Although statistically significant, the observable behaviors were less strongly related to crew-level CRM and technical performance ratings than expected (overall R^2 values were .145 and .133, respectively). As expected, however, crew-level CRM and technical performance ratings were moderately related to individual PIC and SIC ratings, (overall R^2 values were .455 and .442, respectively). Summary results for the six path analyses are presented in Table 8.

Insert Table 8 about here

Mediational tests were performed using Baran and Kenny's (1986) three-step procedure. First, each criterion variable (either PIC or SIC ratings) was regressed on the independent variables (observable behaviors). Next, each criterion variable was regressed on the mediator variables (CRM and technical performance ratings). Finally, the criterion variables were regressed on both the mediator and independent variables entered as a block.

As each Event Set has two dependent variables (CRM and technical performance), twelve separate mediational tests were performed. In general, these tests suggested the presence of fully mediated relationships, as the effect of observable behaviors on PIC and SIC ratings occurred largely through the intervening CRM and technical ratings (Baran & Kenny, 1986). Although the observable behaviors did exhibit statistically significant increments in predictive validity, they were all less than one percent of incremental variance. The high degree of statistical power for these analyses, coupled with the small incremental effects for observable behaviors, suggest that these relationships can be considered fully-mediated.

Post Hoc Tests. For administrative reasons, the LOE at this particular carrier was specifically designed to emphasize the assessment of CRM proficiency, as opposed to technical proficiency. This is due to the fact that technical proficiency is assessed during a previous day of qualification training in the form of a Maneuver Validation. During maneuver validation, crews are required to perform a number of complex safety maneuvers in the simulator, and are assessed as individuals for their technical performance.

Therefore, we decided to test the relative efficacy of CRM and technical proficiency ratings as predictors of PIC and SIC ratings. For each event set, we tested for significant differences between the beta weights of CRM and technical proficiency using equations developed by Cohen & Cohen (1983). Of the twelve statistical tests, six revealed statistically significant results, with greater weight typically being afforded to CRM proficiency ratings (see Table 9).

Insert Table 9 about here

In general, the results suggest that instructor/evaluators weight CRM proficiency higher than technical proficiency in their overall ratings of PIC (.406 vs. .289, respectively) and SIC performance (.406 vs. .276, respectively).

Discussion

The results were mixed. Some of the results were disheartening, such as the factor structure of the observable behaviors. For example, subject matter experts had specifically developed the LOE such that all observable behaviors were linked to specific knowledges, skills, and abilities (KSAs) required for effective performance. Therefore, given the relatively large number of KSAs, the un-interpretable nature of the factor analysis was unexpected.

The factor analytic results could have turned out many ways: a "general" factor could have emerged, a simplex matrix based on phase of flight could have emerged, a simplex matrix based on behavioral requirements could have emerged, etc. However, none of these occurred. Rather, a virtually un-interpretable solution was discovered, regardless of the statistical technique used. The observed results are not inconsistent with the hypothesis that the instructor/evaluators were incapable of distinguishing among the observable behaviors, regardless of the phase of flight and/or the specific KSA under consideration.

Other results were more promising. For example, the results of the two other factor analyses (CRM/technical performance, and PIC/SIC ratings) were not unexpected. Although no *a priori* rationale was provided as to why this particular configuration would emerge across the two factor analyses, it was hypothesized that CRM/technical (and PIC/SIC) ratings from different event sets would be related to one another. Quite simply, this is due to the fact that performance during early phases of flight was expected to influence performance during later phases of flight. Also comforting was the remarkable similarity between the factor structures of the CRM/Technical ratings and the PIC/SIC ratings.

The results of the path analyses were, however, mixed. While the moderately strong linkages between crew-level CRM and technical performance and overall PIC/SIC ratings were encouraging, the relationships between the observable behaviors and CRM/technical performance were somewhat lower than expected. This is especially troubling given the fact that this rating process was specifically designed to establish a link between specific, behavioral indices and crew-level ratings of CRM performance (e.g., to ensure union acceptance of the evaluation technique).

Armed with the observed results, interviews were scheduled with key informants among the population of Boeing 757 instructor/evaluators. Based on the interview data, as well as anecdotal information derived from written comments on LOE evaluation forms, Quality Assurance personnel decided to replace the observable behavior ratings with fewer, more general, topic-level ratings. These topic ratings were designed to occur at a more "natural" level of concept formation, and as such, be less cognitively demanding on the instructor/evaluators (Rosch, 1975). It was hypothesized that the reduced cognitive workload would result in more accurate ratings, and in doing so, would establish a stronger link between observable behaviors and crew-level ratings of CRM and technical performance. As this carrier re-designs it's LOEs on a yearly basis, the switch from observable behaviors to topic-level ratings was included in the new LOE (herein referred to as LOE Bravo).

Method

The second study is a conceptual replication and extension of the first. Therefore, the methodology is essentially the same as for the previous LOE. As before, evaluations were made in the same Boeing 757 simulators under similar conditions. Furthermore, the evaluation process remained the same: while the crew members assumed their normal flight roles, the instructor/evaluators manipulated the simulator, interacted with the crew, and made their evaluations in real time.

There are, however, three notable exceptions. First, the LOE content was changed. This was done to avoid "testing effects" (Cook & Campbell, 1979), and to comply with regulatory requirements which prohibit using the same LOE during two successive years (Federal Aviation Administration, 1990). Second, observable behavior ratings were replaced with fewer, more general, topic-level ratings. Finally, a new sample of 837 crews participated in the study. As stated earlier, however, operational and regulatory constraints render it possible that the subjects from these two studies are not completely independent.

Results

Prior to testing the congruence between the actual ratings and the corporate standard operating procedure, preliminary data checks were again made via examination of the descriptive statistics. The relatively high means and large standard deviations (as can be seen in Tables 10 through 15) suggest that all variables were negatively skewed, although somewhat less than in the first sample (LOE Alpha). This range restriction should attenuate the observed correlations and measures of variance explained. Nevertheless, the larger variances should allow for greater covariation among the observed performance measures than in the previous sample. It should also be noted that carrier personnel had also addressed the problem of missing data. Sample sizes for the following analyses typically averaged around 800 crews.

Insert Tables 10 through 15 about here

Factor Structure. Principal components analyses were performed to better understand the underlying data structure. First, the twenty topic-level ratings (across all six event sets) were factor analyzed using principal components analysis with a varimax rotation. Preliminary data quality checks suggest that the correlation matrix contained a substantial number of non-zero correlations, and as such, was amenable to exploratory factor analyses. For example, the Kaiser-Meyer-Olkin measure of sampling adequacy yielded a value of .864, and Bartlett's test of sphericity yielded a value of 16208.104 (df = 190, $\underline{p} < .000$). Both results fall within appropriate numerical ranges as suggested by Tabachnick & Fidell (1996). A principal components analysis suggested the presence of six components which account for approximately 71 percent of the observed variance. These principal components rotated to simple structure, with the topic-level ratings from each event set loading on their own unique factors. Only one topic-level rating had a single cross-loading greater than .30.

Next, principal components analyses were performed on the twelve crew-level CRM and technical performance ratings (across the six event sets). Again, preliminary data quality checks suggest that the correlation matrix was amenable to factoring. The Kaiser-Meyer-Olkin measure of sampling adequacy yielded a value of .708, and Bartlett's test of sphericity yielded a value of 11595.342 (df = 66, p < .000). Again, these values fall within acceptable mathematical ranges. Four components emerged, accounting for approximately 73 percent of the observed variance. In general, simple structure was found; only two items had a cross-loading in excess of .30. The first component included the CRM and technical ratings from the "Top of Descent to Final Approach" and "Final Approach to Taxi-In" event sets. The second component included the CRM and technical ratings from the "Pre-Departure to Taxi-Out" and "Takeoff to Top of Climb" event sets. The third component included the CRM and technical ratings for the "Reaching Top of Climb to FL280" event set. The fourth and final component included the CRM and technical ratings for "Takeoff to Top of Climb" and "Dangerous Goods Incident" event sets.

Finally, principal components analyses were performed on the twelve PIC and SIC (across the six event sets). Again, the correlation matrix exhibited desirable statistical properties, suggesting that it was amenable to exploratory factor analyses. The Kaiser-Meyer-Olkin measure of sampling adequacy yielded a value of .680, and Bartlett's test of sphericity yielded a value of 13838.224 (df = 66, p < .000). As before, the solution rotated to simple structure, with only two items exhibiting cross-loadings in excess of .30. However, the factor structure which emerged was somewhat different from the one on the CRM and technical performance ratings. Five principal components emerged. These five components account for approximately 83 percent of the observed variance. Like the previous analysis, the first component included the PIC and SIC ratings from the "Top of Descent to Final Approach" and "Final Approach to Taxi In" event sets. The second component included the PIC and SIC ratings from the "Final Approach to Taxi In" and "Dangerous Goods Incident" event sets. The third component was composed of the PIC and SIC ratings for the "Reaching Top of Climb to FL280" event set. The fourth component was comprised of PIC and SIC ratings on the "Takeoff to Top of Climb" event set. The fifth and final component was composed of the PIC and SIC ratings for the "Pre-Departure to Taxi-Out" event set.

Path Analyses. Finally, the instructor/evaluators' rating process was compared to the carrier's standard operating procedures using path analytic techniques. Again, the path analyses were performed according to specifications set forth by Cohen & Cohen (1983) and Pedhazur (1982). As in LOE Alpha, the topic-level ratings were only moderately related to crew-level CRM and technical performance ratings (overall R^2 values were .165 and .169, respectively). However, crew-level CRM and technical performance ratings were much more strongly related to individual PIC and SIC ratings, with R^2 values being .624 and .561, respectively. Summary results for the six path analyses are presented in Table 16.

Insert Table 16 about here

Finally, mediational tests were performed using Baran and Kenny's (1986) three-step procedure. First, each criterion variable (either PIC or SIC ratings) was regressed on the independent variables (topic-level ratings). Next, each criterion variable was regressed on the mediator variables (CRM and technical performance ratings). Finally, the criterion variables were regressed on both the mediator and independent variables entered as a block.

As each Event Set has two dependent variables (CRM and technical performance), twelve separate mediational tests were performed. In general, these tests suggested the presence of fully mediated relationships, as the effect of topic-level ratings on PIC and SIC ratings occurred largely through the intervening CRM and technical ratings (Baran & Kenny, 1986). Although the topic-level ratings did exhibit statistically significant increments in predictive validity, they were all less than one percent of additional variance. Given the high degree of statistical power available, and the small incremental effects, these relationships can be considered essentially fully-mediated.

Post Hoc Tests. Again, the relative efficacy of CRM and technical proficiency of PIC and SIC performance ratings was assessed. In general, the results suggest that instructor/evaluators weight CRM proficiency much higher than technical proficiency in their overall ratings of PIC (.451 vs. .325, respectively) and SIC performance (.459 vs. .272, respectively), with nine of the twelve tests reaching statistical significance (see Table 17 for more information).

Insert Table 17 about here

Practically speaking, the results provide converging evidence to support the usefulness of the LOE evaluation procedure. Specifically, the LOE was designed to assess CRM, as opposed to technical proficiency, and appears to do just that.

Discussion

The results from the LOE Bravo were quite different from LOE Alpha. For example, unlike the observable behaviors, which emerged as ten uninterpretable factors, the topic-level ratings cleanly emerged as six separate factors (by their respective event sets). This is encouraging, as the LOE Bravo was specifically designed to minimize the instructor/evaluators' cognitive workload[2]. The observed factor structure of the topic-level ratings indicrectly suggests that instructor/evaluators can distinguish among behavioral indices (e.g., topic-level ratings) of CRM performance. In this case, the instructor/evaluators were apparently making their distinctions based on phase of flight.

The factor structures for the CRM/Technical performance ratings and PIC/SIC ratings, however, were somewhat unexpected. Unlike, LOE Alpha, in which the two factor structures were identical, the factor structures for LOE Bravo were somewhat different. The implications of this remain unclear; although current efforts are underway to identify causes for these relationships.

Most importantly, however, the path analytic results suggest that instructor/evaluators were able to link behavioral indices with overall, crew-level evaluations of CRM performance. Furthermore, the effect of behavioral indices (topic-level ratings) on summary ratings (PIC and SIC) occurred entirely through the effect of crew-level CRM and technical performance ratings. The results of the mediated tests (from both LOEs) suggest that the instructor/evaluators were using the specified evaluation process as it was designed.

Like all studies, however, this one asks more questions than it answers. First, while the linkage between concrete behavioral examples and CRM/technical performance was both practically and statistically significant, they are still much lower than expected. Given that both observable behaviors and topic-level ratings account for only about 15 percent of the observed variance in CRM and technical performance, the logical question is "what is accounting for the other 85 percent?". Unfortunately, the only answers that we can offer are speculative.

First, it must be noted that the behavioral examples were never meant to be an exhaustive list. Therefore, there may be additional topics/behaviors (not listed on the evaluation worksheet) that account for the "error" variance. Follow-up analyses, such as focus groups meetings between instructor/evaluators and LOE developers may shed some light on this hypothesis.

Second, even though the topic-level ratings exhibited higher variances than their observable behavior counterparts, a substantial degree of range restriction still remained, which may serve to depress the covariances among the measures. Given operational (e.g., the physical conditions under which ratings were made) and administrative constraints (e.g., this was a "jeopardy" evaluation of maximal performance) the observed results may very well represent a "ceiling effect" between behavioral indices and crew-level evaluations. This would not be uncommon, as the performance appraisal literature has consitently shown that it is extremely difficult to obtain accurate performance ratings (Murphy & Cleveland, 1995). Again, follow-up analyses must be performed to directly test this hypothesis. If range restriction does not occur under non-jeopardy conditions, such as a Line Oriented Flight Training (LOFT), when presumably crews are performing under typical levels of motivation, this would provide corroborating evidence for the "administrative use" hypothesis.

Nevertheless, the observed results are encouraging. Specifically, it is encouraging to note that under conditions of (presumably) lower cognitive workload, instructor/evaluators' ratings of crew performance exhibited a more diversified factor analytic structure, and stronger path analytic results were observed in the rating process. However, we do recognize the need to replicate these results with different fleets, different evaluators, and different LOE content to ensure external generalizability.

One question that remains unanswered, however, concerns the nature of the PIC and SIC performance ratings. Although designed to be individual-level ratings, they exhibit some degree of "crewness", as evidenced by the high intra-class correlations. The logical question then becomes "what exactly are they measuring"? At this time, we can offer no definitive answers. Clearly, they are not identical with CRM and technical proficiency ratings; if they were, the R^2 values obtained from the path analyses would have been much higher. This remains a vexing administrative issue, especially if such ratings are used for human resource purposes such as promotion/disciplinary actions.

Finally, the LOE rating process and the associated statistical techniques discussed herein represent an important advance in the evaluation of CRM performance in LOS environments. By exploring the underlying patterns of ratings, as well as their correspondence to the carrier’s standard operating procedures, both researchers and fleet personnel can be more confident in the results that they obtain. Furthermore, by noting which aspects of the LOE evaluation (statistically) did/did not work as planned, the results also suggest areas for improvement during the development of future LOEs. Therefore, we propose that when evaluating CRM performance data, these analyses be performed as initial "data checks" prior to performing other multivariate tests of statistical inference.

Conclusions and Recommendations

At the beginning of this paper, we noted that the existing body of CRM training evaluation research is deficient. For example, it has been suggested that there is reason to believe that the relationship between attitudes and measures of typical job performance is relatively weak (Alliger et al. 1997). Likewise, evaluation studies which rely on measures of maximal performance, such as LOE ratings, may be subject to a number of serious statistical errors, such as range restriction. So, what are we to make of all this?

Let us begin by reminding the reader that CRM research is a relatively new field of inquiry. Like all fields of research during their "adolescence", CRM training and evaluation will experience a number of growing pains. Fortunately, signs of improvement are on the horizon. For example, recent work by Holt and colleagues (George Mason University, 1996; Williams et al, under review; Holt et al., 1996) have provided methods such as Inter-Rater Reliability (IRR) training to assist instructor/evaluators in making ratings which are not only consistent with one another, but are also sensitive to differences between crews of varying performance levels. Current projects by Johnson, Goldsmith and colleagues (personal communication) and Baker and colleagues (personal communication) are extending this work by attempting to train evaluators to a "gold standard" of accuracy (e.g., referent rater reliability or RRR). Although we recognize the limitations of psychometric theory to performance evaluation (Murphy & Cleveland, 1995), we believe that the integration of these two lines of rater training has the potential to enhance the accuracy of future evaluation efforts. At the same time, future IRR and RRR research should incorporate the multi-dimensional nature of rating accuracy in their analyses and training procedures (Chronbach, 1955; Murphy & Balzer, 1989).

Similarly, work by Holt et al. (1998) has shown the promise of using "triangulation" procedures to make up for the deficiencies of individual methodological techniques. While we recognize the increased cost of doing so, we encourage other researchers to adopt similar multi-method evaluation designs. Quite simply, we believe that such techniques represent the surest way to determine whether or not CRM training programs are, in fact, meeting organizational and regulatory criteria for effectiveness. Further, we believe that until such techniques become the norm, CRM research will be harshly criticized on methodological grounds.

Finally, we are pleased to notice the re-introduction of well-articulated theoretical models into the CRM literature (Helmreich, 1998; Helmreich & Merritt, 1998). This is a long overdue addition to the field. Much like the social psychological research of the 1950's and 1960's, we fear that without integrative theory, CRM research will proceed along stagnant lines of restrictive research. Above all, we feel that the introduction of well-articulated theoretical models will do much to dispel the notion that CRM training is just "another management fad".

In summary, there is much to be critical about the current state of CRM research. At the same time, there is also much promise. The introduction of new rater training techniques, the use of "triangulation" designs, and the re-emergence of theory are all causes for celebration.

Notes

Note #1: All exploratory factor analyses were performed using principal components analysis with varimax rotation. While a number of different extraction and rotation techniques were performed, all yielded remarkably similar results. Furthermore, all solutions required fewer than 20 iterations to converge, thereby suggesting mathematically stable results.

Note #2: Given operational constraints, a direct test of the cognitive workload hypothesis was not possible. However, indirect support for this conclusion is provided by interviews with the instructor/evaluators, archival data derived from LOE evaluation forms, and the marked changes in both factor structures and estimates of effect size.

References

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. Personnel Psychology, 50(2), 341-358.

Baran, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173-1182.

Boeing Commercial Aircraft Group. (1994). Statistical summary of commercial jet aircraft incidents, worldwide operations 1959-1994. Seattle, WA: Author.

Chronbach, L. J. (1955). Processes affecting scoreson understanding of others' and "assumed similarity." Psychological Bulletin, 59, 177-193.

Clothier, C. C. (1991). Cockpit resource management: Effects of behavioral interactions across airlines and aircraft types. Unpublished master's thesis, The University of Texas at Austin.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Earlbaum.

Cohen, J. & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Earlbaum.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Boston, MA: Houghton Mifflin.

Dubois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximal performance criteria: Definitional issues, prediction, and White-Black differences. Journal of Applied Psychology, 78(2), 205-211.

Federal Aviation Administration. (1993). Crew resource management training. Advisory Circular #120-51A. Washington, DC: Author.

Federal Aviation Administration. (1990). Line operational simulations: Line oriented flight training, special purpose operational training, line oriented evaluation. Advisory Circular #120-35B. Washington, DC: Author.

George Mason University. (1996, February). Improving crew assessments. Manual to accompany the inter-rater reliability (IRR) workshop sponsored by the Federal Aviation Administration's Office of Aviation Research: Fairfax, VA: Author.

Gregorich, S. E., & Wilhelm, J. A. (1993). Crew resource management training assessment. In E. L. Weiner, B. G. Kanki, & R. L. Helmreich (Eds.), Cockpit resource management (pp. 173-198). San Diego, CA: Academic Press.

Helmreich, R. L. (1998, September). The downside of having a brain: Reflections on human error and CRM. Presentation made at the ATA CRM Industry Workshop.

Helmreich, R. L. (1987). Exploring flight crew behavior. Social Behaviour, 2, 63-72.

Helmreich, R. L. (1984). Cockpit management attitudes. Human Factors, 26(5), 583-589.

Helmreich, R. L., & Merritt, A. C. (1998). Error and error management. University of Texas Aerospace Crew Research Project. Technical Report #98-003. Austin, TX.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (in press). The evolution of crew resource management training in commercial aviation. International Journal of Aviation Psychology.

Helmreich, R. L., & Wilhelm, J. A. (1991). Outcomes of crew resource management training. International Journal of Aviation Psychology, 1(4), 287-300.

Helmreich, R. L., & Foushee, H. C. (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E. L. Weiner, B. G. Kanki, and R. L. Helmreich (Eds.), Cockpit resource management (pp. 3-45). San Diego, CA: Academic Press.

Holt, R. W., Seamster, D. A., Hansberger, J. T., Beaubien, J. M., & Incalcaterra, K. (1998). Evaluation of proceduralized CRM training at a regional airline. Unpublished manuscript. Fairfax, VA: George Mason University.

Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1996). Application of psychometrics to the calibration of air carrier evaluators. Unpublished manuscript. Fairfax, VA: George Mason University.

Howell, D. C. (1994). Statistical methods for psychology (4th edition). Belmont, CA: Duxbury Press.

Kanki, B. G., & Palmer, M. T. (1993). Communication and crew resource management. In E. L. Weiner, B. G. Kanki, & R. L. Helmreich (Eds.), Cockpit resource management (pp. 99-136). San Diego, CA: Academic Press.

Kozlowski, S. W. J., Chao, G. T., & Morrison, R. F. (1998). Games raters play: Politics, strategies, and impression management in performance appraisal. In J. W. Smither (Ed.), Performance appraisal: State of the art in practice (pp. 163-205). San Francisco, CA: Jossey-Bass.

Lauber, J. K. (1984). Resource management in the cockpit. Air Line Pilot, 53, 20-23.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. Journal of Applied Psychology, 74(4), 619-624.

Murphy, K. R., & Cleveland, J. N. (1995). Understanding performance appraisal: Social, organizational and goal-based perspectives. Thousand Oaks, CA: Sage.

National Transportation Safety Board. (1994). A review of flightcrew-involved major accidents of U. S. air carriers, 1978 through 1990. Safety Study NTSB / SS-94 / 01, Notation 6241. Washington, DC: Author.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: Explanation and prediction (2nd edition). Fort Worth, TX: Harcourt Brace.

Prince, C., Oser, R., Salas, E., & Woodruff, W. (1993). Increasing hits and reducing misses in CRM/LOS scenarios: Guidelines for simulator scenario development. International Journal of Aviation Psychology, 3(1), 69-82.

Rosch, E. H. (1975). Cognitive representation of semantic categories. Journal of Experimental Psychology: General, 104, 192-233.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximal performance. Journal of Applied Psychology, 73(3), 482-486.

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, Office of the Chief Scientific and Technical Advisor for Human Factors, AAR-100.

Tabachnick, B. G., & Fidell, L. S. (1996). Using multivariate statistics (3rd edition). New York: Harper Collins.

Williams, D. M., Holt, R. W., & Boehm-Davis, D. A. (under review). Training statistical skills to non-statisticians: A case study of inter-rater reliability training for pilot instructor/evaluators. International Journals of Training and Development.

Table 1

A Comparison of Topic-Level Ratings and Observable Behaviors (Matched for Event Set Content)

LOE Alpha (Topic-Level Ratings)	LOE Bravo (Observable Behaviors)
Pre-Departure to Taxi Out	Pre-Departure
Captain ensures pre-departure tasks are complete. Crew members advocate correct and relevant plan. Crew recognizes clearance change on the PD. Crew members contribute information concerning all…	Captain and crew discuss the effects the forecast… Each crew member contributes information concerning all aspects of the operation. Crew discussion about the turbulence includes contingencies. Cockpit environment exists for all crew members to openly express ideas. Captain ensures all pre-departure tasks are completed per SOP.
Take-Off to Top of Climb	Taxi Out
Runway change is discussed with each crew member. Crew members take initiative to complete tasks necessary… Information is transmitted with sufficient time.	Runway change is discussed/performance data is reviewed to determine…. Information is contributed by all crew members and decision to accept or reject… Crew members take initiative to complete all tasks and ensure the… If overloaded, crew members bring to attention of the rest of the crew.
Top of Climb to FL 280	Climb
Crew articulates increased fuel burn and sets… monitors current fuel state.	Crew correctly copies and reads back the route change to San… Before deciding to accept the new clearance, crews review the new routing. The PNF enters the route change, if accepted into the FMC and crew members… When new information is received, it is used to critique previous decisions.

Note. Ellipses are used to indicate additional text information (which is unavailable) due to character limitations in the Microsoft Access database from which they were extracted.

Table 1 (Continued)

LOE Alpha (Topic-Level Ratings)	LOE Bravo (Observable Behaviors)
Dangerous Goods Incident	Cruise
Crew acknowledges flight attendants report. Captain uses all information gathered. Environment exists for crew members to freely… Crew correctly identifies problem, advocates…	Crew recognizes the EICAS message and correctly interprets its meaning. Cockpit environment exists for crew members to openly express ideas. After discussion with the FO, ATC, Dispatch, and SAM, Captain considers… Crew members maintain awareness of workload conditions in themselves. Operational decisions are clearly stated to all affected parties. Crew discusses implications the engine configuration has on it's final approach. When new information is received, it is used to critique previous decision.
Top of Descent to Final Approach	Descent
Crew members openly express concerns. Crew members verify all messages between members. Crew members continue to advocate plans with clear… Continued use of critique provides crew information.	Crew verbalizes recognition of the increasing oil temperature. Crew recognizes implications a single engine has and reviews previous… Operational plans and decisions are clearly stated and other crew members Crew divides cockpit into PF and PNF duties. Each crew member monitors. Crew sets bottom lines along diversion plans to monitor their progress Crew members assume tasks to ensure the aircraft is ready for the approach.
Final Approach to Taxi-In	Approach
Crew members use previously-defined points and so… Crew members take initiative to complete all tasks. Crew members continue to communicate task priorities.	Crew flies the approach and landing as previously briefed. Crew continues to monitor their progress towards previously established goals. Crew conducts extensive approach briefing. Captain assumes all procedures… Crew members take initiative to complete all tasks necessary to ensure… Crew members clearly communicate ideas, opinions, and tactical decisions. If unable to accept clearance, crew immediately requests…

Figure Caption

Figure 1. Graphical depiction of the carrier’s standard operating procedure (SOP) for making individual- and crew-level performance ratings during an Event Set.