

## **Improving Rater Calibration in Aviation: A Case Study**

Robert W. Holt, Jeffrey T. Hansberger, and Deborah A. Boehm-Davis

George Mason University  
Psychology Department  
MSN 2E5  
Fairfax, VA 22030-4444  
USA

Running Head: Improving Rater Calibration

**To appear in: International Journal of Aviation Psychology**

### Abstract

The assessment of pilot performance relies on systematic observation and assessment by a single trained rater or instructor/evaluator (I/E). Due to the importance of aviation safety, it is imperative that the rating and evaluation processes used by these I/Es yield reliable and valid data. This paper describes a case study focused on improving the reliability and validity of crew assessment at a commercial carrier. The process for improving reliability involved the evaluation of current training methods, the construction and evaluation of five metrics for assessing inter-rater reliability, and a standardized process for using these metrics to train I/Es. A separate set of interventions was developed and implemented to improve structural validity. Data collected from two fleets in this airline over a three-year period suggest aspects of reliability and validity that should be the focal points of continuing I/E training.

## **Improving Rater Calibration and Performance in Aviation**

Robert W. Holt, Jeffrey T. Hansberger, and Deborah A. Boehm-Davis  
George Mason University

Pilot performance, as assessed by Line Operational Flight Training (LOFT), Line Operational Evaluation (LOE), Maneuver Validation, Line Checks, or Proficiency Checks, relies on systematic observation and assessment by a single trained rater or instructor/evaluator (I/E). This performance information is used to determine the effectiveness of training and overall safety levels for a fleet. Therefore, airline safety and effective training ultimately depend on I/E assessments. The role of the I/E as a rater for crew certification is critical in this process.

It is imperative that the rating and evaluation process of the I/Es yields reliable and valid data. Evaluators should be reliable both in the stability of their rating criteria over time and their consistency with the rest of the I/E population. The evaluation process should also be valid in that the evaluators are rating what they are supposed to be rating. That is, ratings of key components of performance should not be affected by extraneous factors. This paper describes a case study that focused on developing methods to index measurement reliability and validity of I/E judgments at a commercial carrier and institute systematic I/E training based on this information.

The initial focus of this research project was an operational evaluation of improved Crew Resource Management (CRM) training called Advanced Crew Resource Management (ACRM). This training was implemented and tested over a three-year period at a regional air carrier (see Boehm-Davis, Holt, & Seamster, in press; George Mason University, 1996 for more information). Very early in this research project, both carrier training personnel and the research team recognized a need to establish the current levels of reliability and validity for I/E evaluations in order to have some confidence in the scientific quality of the training evaluation data. Adequate reliability and validity were a prerequisite to having an accurate evaluation of any improved crew training resulting from the implementation of ACRM.

Initial work with the carrier suggested that elements of rater reliability and validity should be addressed. For reliability, the carrier did not have any way to systematically evaluate ratings or to determine if the I/Es were rating the pilot population consistently. To help alleviate these potential errors, Birnback & Longridge (1993) advise that airlines train their I/E population. Borman (1975) has shown that rater training can be effective. Aside from carrier trainers and management, the pilots and the union also agreed that ensuring reliable ratings was important for fair and unbiased evaluations. Validity was a concern because initial development and implementation of the new CRM program uncovered a wide range of different opinions among the I/Es on how to define and proceduralize "good" CRM. Secondly, the development of Line Operational Evaluation (LOE) flight scenarios designed to evaluate crew technical and CRM performance (Hamman, Seamster, Smith, & Lofaro, 1991) also surfaced different opinions about the specific crew behaviors that would demonstrate competent performance at the end of training. Given these issues, the research team reviewed common reliability metrics and ways to improve validity.

### Reliability

The investigation of existing approaches for improving reliability and validity revealed indexes and techniques primarily designed to improve reliability. This may be due to the psychometric viewpoint that reliability is a precondition for validity (Nunnally, 1967). It may also be due to the difficulty in obtaining a precise index of validity. The review also indicated two rather distinct approaches to training reliability. The first focused on consistent co-variation of raters across items (McGraw & Wong, 1996; Jako & Murphy, 1991; Viswesvaran, Ones & Schmidt, 1997). One common index was the Pearson product-moment correlation among raters across rated items. The critical idea is that when rating a complex stimulus across different items, raters should consistently shift to either higher or lower evaluations on each item (indicated by high values for inter-rater correlations). This idea was adapted for this project by using the

average inter-rater correlation as a *consistency* index (see George Mason University, 1996 for a description of the metrics and their visualizations).

The second distinct approach emphasized the agreement of a set of raters on a particular evaluation item. One common index was the agreement index ( $r_{wg}$ ) that compares the variance of the distribution of ratings for a specific item with the variance of a flat or uniform distribution of ratings for that item (James, Demaree, and Wolf, 1984, 1993; Law & Sherman, 1995). The critical idea is that raters should have a high degree of agreement in their ratings for each item (indicated by low variance and a value of  $r_{wg}$  near 1.0). This idea was adapted for this project by using the  $r_{wg}$  as an *agreement* index and using low values of the index to trigger discussion of rating differences on each item.

Although these indexes gave valuable information about different aspects of rater reliability, they were insufficient. In particular, three critical issues that directly or indirectly affect rater reliability were not covered in a manner that could be directly communicated to the I/Es. This situation seemed to call for a multi-component approach to rater reliability such as that advocated by Saal, Downey, & Lahey (1980). The first issue was related to the overall distribution of ratings given by a rater for a set of stimuli. If individual raters were using the rating scales differently, these differences might result in different distributions of ratings. Using an example of a 4-point rating scale, some raters might use extreme values (1 and 4) while others might use central values (2 and 3).

Conceptually, any differences in the shape of rating distributions would limit the possible values of both the inter-rater correlations and the average  $r_{wg}$  index across the set of items. Practically, differences in rating distributions would signal undesirable differences in the use or interpretation of the rating scale or observed stimuli. Therefore, a measure of distribution similarity, labeled *congruency*, was developed to indicate the degree to which each I/E's distribution of ratings matches the group's distribution of ratings. A rater whose distribution of

ratings matches the group would be high on the congruency index whereas a rater with a very different distribution of ratings would be low.

The second issue was related to the problem of raters who give more lenient or more strict ratings than the group as a whole. The I/E cadre and pilots at the carrier were not only aware of this problem but had even given specific labels to the raters who were characteristically lenient (a Santa Claus rater) or characteristically harsh (an Axe Man rater). Conceptually, this difference in average ratings would not necessarily affect the inter-rater correlation at all. Similarly, although these systematic differences would lower average values of  $r_{wg}$ , the presence of low values for  $r_{wg}$  would NOT necessarily indicate systematic differences. Therefore, a method called *systematic differences* was developed to reflect this specific problem. The analysis for systematic differences uses a *t*-test index to compare each I/E's average rating to the group's average rating. This comparison identifies the I/Es who rate significantly higher or lower than the group.

The third issue was the inability of raters to correctly and consistently make fine-grained discriminations of crew performance. In particular, the ability to discriminate unsatisfactory from barely satisfactory performance, and satisfactory (FAA minimum) from company standard performance were considered critical. A common problem at this carrier was the inability of the I/Es to discriminate among these performance levels.

To counteract this tendency, a final index, based heavily on the frame-of-reference training technique (Bernadin & Buckley, 1981), was developed. This index required a group of SMEs to establish the actual levels of crew performance for a videotaped segment of performance. Using the actual SME ratings, Hay's omega-squared strength of effect measure is calculated to indicate how well each I/E was changing his or her average rating relative to the group of SMEs "true rating". This index was labeled *sensitivity*.

An initial small-group tryout of an initial version of the I/E training gave promising preliminary results (George Mason University, 1996). Over a 2-month period, a test group of five raters showed some improvement. They reduced systematic differences from an average

difference of .20 to .18 among the raters; increased consistency from an average inter-rater correlation of .56 to .63; and increased agreement on the four items that were the focus of group discussion from .14 to .85. Having developed this approach for assessing reliability, we also considered methods for assessing the validity of assessment.

### Validity

For this project, a form of construct validity that emphasized structural validity was used for validity assessment. The structural validity assessment focused on the expected relationship among the different ratings in the performance evaluation. In particular, we used structural validity techniques to assess the validity of the LOE rating process. The I/Es were trained in the systematic use of evaluation sheets for each segment or event set of the LOE. For each segment, I/Es first rated specific observable behaviors; they then rated intermediate-level technical and CRM skills. Finally, they assigned overall evaluations of Pilot in Command (PIC), Second in Command (SIC) and crew performance. To determine the structural validity of this process, a path analysis was performed to determine the relationships among specific observable behaviors, intermediate technical and CRM skill ratings, and PIC, SIC and crew performance ratings. For simplicity and robustness, multiple regression was adopted as the basis of the path analysis (Cohen & Cohen, 1983). Data from the structural validity analysis in the form of path diagrams were used to inform the I/Es of validity issues in separate sessions.

### I/E Rater Training

Since reliability is a precondition to validity (Nunnally, 1967), our I/E rater training focused on the reliability indexes. We combined the set of reliability indexes into a training package for the I/Es that did not mimic any specific training approach but rather was based on basic training principles. The training included the traditional metrics of inter-rater correlation and agreement together with the congruency, systematic differences, and sensitivity indexes. For each index, a visualization was developed so that the information could be easily presented for feedback to I/Es

(Holt, Meiman, & Seamster, 1996). The two approaches to rater training that were adapted to this context were rater error training and frame of reference training.

Although Woehr & Huffcutt (1994) found that rater error training is not the most effective rater training method, it seemed necessary for I/Es to see how they were making rating errors compared with the group. Since each I/E at this carrier was an experienced pilot, there was a tendency to believe that each one had the one, true, and correct evaluation of crew performance. This cognitive set had to be broken before the constructive work could be done to change the rating standards and processes. Therefore, one component of the training was individual feedback for each I/E (Baker, & Mulqueen, 1999) on how they were different from the group. This includes feedback on his or her average correlation with the other raters, congruency of rating distribution with the class, systematic differences from the average evaluation, and sensitivity to differences in crew performance as determined by SMEs.

Frame-of-reference training has been shown to be one of the more effective methods for training raters (Woehr & Huffcutt, 1994). For this reason, establishing a common frame-of-reference for ratings was pursued in two ways. First, each training event included several hours of detailed discussion of the justification for ratings on the low-agreement items. The goal of this discussion was finding rating principles that would cause all raters to come to a common rating on the low-agreement items. This aspect of the training used principles developed by the group as it achieved consensus on low agreement items to form a frame-of-reference for ratings. Since no *a priori* external frame-of-reference such as a gold standard (Baker & Dismukes, in press) was available, the principles derived from the consensus process were used to construct a common evaluation framework. Secondly, as discussed before, the sensitivity index provides a more traditional frame-of-reference provided by an external group of SMEs.

The final rater training strategy included is consistent with performance dimension training (Woehr & Huffcutt, 1994). The primary purpose of performance dimension training is to ensure that raters know and can use the rating scales for their evaluations. The rating process was

structured by the LOE worksheet that led the rater from observable behaviors to skill ratings to overall evaluations. The structural validity analysis examined how well this process was carried out in LOE evaluations. From these analyses, the path diagrams with standardized regression coefficients for each significant link were used for feedback to the I/Es.

The constraints of the carrier's training program were such that a maximum of 6 hours during one day were available to deliver I/E rater reliability training every 8 months to a year. To avoid overwhelming the I/Es with statistical information, validity feedback was delivered at separate sessions. Compared to training programs such as assessment centers that may involve many days or weeks of training effort, this training program used minimal time and resources.

A second constraint was due to the fact that the cooperating airline was a regional carrier. As is typical for regional carriers, pilots often transition either between fleets or to a larger domestic carrier. Since it is important to maximize flying time to have a chance of being hired by a major carrier, there is a strong disincentive for the I/Es to stay in the training division for any length of time. Consequently, the I/E cadre experienced severe turnover each year during the course of this study.

These constraints must be kept in mind when considering the results of this study. The values for reliability and validity that were achievable in this context are predicated on the limited time, personnel, and turnover conditions of a typical regional carrier.

## METHOD FOR TRAINING RELIABILITY

### Sample

The participants in this study were the cadre of instructor-evaluator (I/E) pilots from two fleets within a regional air carrier. The I/Es from each fleet were trained as a separate group. The results presented here represent these two groups of I/Es consisting of 6-16 I/Es each. These two groups typically participated in one IRR session per year. The I/Es are the same ones who evaluated all pilots on the yearly LOE that provided the data for the validity analyses. There usually is, and was in our study, a large turnover rate among the I/E population within regional air

carriers. For example, in one year of this study, the turnover rate was well over 50%. Thus, the results are shown for different numbers of instructor-evaluators across the three years of this study.

### IRR Training

IRR training was implemented and tracked over a three-year period. Delivery of the training program for improving inter-rater reliability involved: 1) developing the metrics and visualizations for measuring reliability, 2) preparing materials prior to the workshop, 3) delivery of the training program in a workshop setting, and 4) development of post-session summary feedback.

### Metrics for assessing reliability

IRR training relies on the five basic indexes for inter-rater reliability (IRR), as described in the introduction. Visualizations were developed to provide feedback to raters on their performance relative to their peers. Each index assesses an important and required element for IRR and when the indexes are combined, they offer a comprehensive approach to IRR. The five IRR indexes are 1) systematic differences 2) congruency, 3) consistency, 4) sensitivity, and 5) agreement. All of these indexes are focused on the reliability of judgment except sensitivity, which is focused on the ability of the I/E to make accurate and fine discriminations in judgment.

### Preparation of materials

In order to assess reliability, a set of raters must evaluate some aspect of human performance. In this case, reliability was assessed by having the I/Es rate videotaped flight scenarios flown by actual pilots. This required the construction and recording of these flight scenarios prior to the training session. The flight scenarios were recorded in a full-motion simulator during an LOE using pilots who were certified to fly that aircraft type. The yearly recurrent LOEs were designed to serve as a work sample during which maximal performance in both normal and abnormal situations could be assessed (Prince, Oser, Salas, & Woodruff, 1993). The scenarios were chosen by a group of subject matter experts (SMEs) to cover a range of

performance activities and categories. Each scenario chosen for IRR training roughly matched an event set or phase of flight. In general, four to six scenarios provided a sufficient number of items to be rated and used in computing the IRR indexes.

In addition to the scenarios to be rated, the I/Es needed to have available to them the materials used during normal evaluations. This usually required copying existing LOE worksheets used for evaluating each event set and scenario guides from the LOE and organizing them into one package per I/E. The LOE worksheet simplified, organized, and standardized the evaluation process.

Finally, prior to the IRR training session, it was necessary to establish the *a priori* levels for the sensitivity evaluations. These *a priori* levels were used to evaluate how well the I/Es could differentiate different levels of performance across segments. Once the worksheet package was organized, the performance segments were evaluated by a separate group of SMEs, which ranged from one to three SMEs, to establish the actual level of performance for each event set to be evaluated.

#### Delivery of the Training Program

Once the metrics and visualizations were developed, and the background materials prepared, the IRR training program was offered at the regional carrier. The program was designed to be a process that incorporates 1) rater error training; 2) frame of reference training, or the clarification of rating standards and scale usage; and 3) problem solving processes as a methodology to approach IRR deficiencies. The actual IRR training session was composed of three primary components: 1) initial ratings of the videotaped scenarios, 2) analysis of the ratings, and 3) a feedback and discussion session.

Initial I/E evaluations. At the start of the training, the IRR training facilitator presented a few prepared slides on the goal of the training. The facilitator also gave instructions for the I/Es to view and independently rate the videotape segments, making their evaluations as they would in any regular LOFT or LOE. The I/Es were cautioned not to disturb or influence the evaluations of

the other I/Es in any way. Each I/E was asked to write a personal identification number (PIN) or their initials on their LOE worksheets in order to present them with individual feedback.

Although PIN numbers maintain I/E anonymity during the group feedback phase, it has been our experience that I/Es prefer using their initials as they want to be identified with their evaluations during the group discussion.

The primary focus of this phase of IRR training was the I/Es viewing the videotape scenario segments and making their evaluations. Before showing each video segment, the facilitator provided the context and relevant situational parameters of the scenario and answered any contextual questions among the group. After this introduction, the I/Es made their evaluations using the LOE worksheets while they viewed the performance of the crew in the segment. Once the segment was played back, the I/Es were given a few moments to finish the evaluations for that segment and then introduced to the next segment. This process was continued until all the scenarios had been viewed and evaluated. At this time, all the worksheet packets were collected by the facilitator and submitted for data analysis.

Data analysis. Once all the data were collected from the I/Es, the data were entered into a Microsoft Excel spreadsheet and analyzed by the researchers to evaluate the five IRR indexes (Note: the IRR macros and instruction on their use are available from the authors). Once data analysis was complete, individual and group feedback was prepared. Individual feedback consisted of paper copies or a separate electronic file containing the visualizations comparing that individual I/E against the group data. The group feedback consisted of slides showing the summary group visualizations.

Feedback Session. The feedback phase lasted two to four hours, depending on the number and types of problems discovered from the data analysis. The facilitator presented feedback on the group overall result for systematic differences, congruency, consistency, and sensitivity. At a more detailed level, I/Es were also given feedback on how their individual evaluations compared with the group (e.g., are they a “Santa Claus” or “Axeman”?, are they congruent and consistent

with the group?, are they sensitive to different levels of crew performance?) for each IRR index. I/Es were not told that they had to change; rather, the general issues underlying the calibration problems were discussed. For example, after feedback on systematic differences among the raters was presented, the general issues that could cause these differences were explored through a facilitated discussion. For systematic differences, potential causes included differences in the interpretation of the judgment anchors for each step of the rating scale used on the LOE worksheet, or the application of personal standards instead of the defined criteria in making evaluations.

The last index covered in the feedback phase was the agreement index. The facilitator presented each item that showed significant disagreement among the ratings (i.e.,  $r_{wg} < .70$ ). As each item was presented, the I/E group was prompted to discuss any aspects that might cause the disagreement among their ratings. A sequence of prompting questions such as, “Did you see the same thing?”, “Did you interpret the behaviors the same way?”, and “Did you judge the behavior using the same criteria?” were effective in guiding the discussion because they tracked the major stages of the observation and evaluation process. These probe questions sparked discussion about both the rating process and the frame of reference for evaluation among the I/Es.

#### Post-session Summary Feedback

The final stage in the IRR training process was to document the results for future reference and distribute them at the organizational level. Particular aspects of the results were directed to the appropriate people and departments after the training session for follow-up work and revisions. For example, if the qualification standards for a procedure were ambiguous, the training department would develop clearer performance standards. These documented results consisted of the 1) quantitative results of the IRR training session and 2) qualitative results (i.e., the topics discussed and the results of the rating process discussions conducted by the I/Es).

### Validity Training

During an evaluation such as an LOE, I/Es make many single-item evaluations and it is important that these evaluations be made in a valid manner. Valid evaluations are evaluations that are reliably made, address the intended behaviors and skills, and are based on the criteria or process the I/Es were trained to use. Our assessment of validity focused primarily on the last topic, how well the I/Es were using the process and criteria they were trained and instructed to use (i.e., structural validity).

For this research project, initial validity training was accomplished as the LOE for each year's evaluations was implemented, which was prior to the IRR training. This training presented results from their past fleet LOE evaluations and reemphasized the evaluation structure and process they learned in their initial I/E training. In this training, I/Es were taught to make their evaluations following worksheets that specified the evaluation scale and evaluation sequence for each event set of the LOE. This section describes 1) the development of the standard rating scale used on the worksheets, 2) the development of the LOE worksheet used to facilitate the evaluation process, and 3) the process and sequence I/Es were trained to use for their evaluations.

#### Standard Rating Scale

One critical element for validity training was the development of a standard rating scale. Using a standard rating scale reduced the training time required to familiarize I/Es with different assessment instruments (e.g. LOE, LOFT, Line Check). Using a standard scale for all forms of assessment effectively increased the amount of practice that I/Es had with the scale. Increased practice should lead to better assessment skills and ultimately better rater reliability and validity, especially given the complex tasks and high workload faced by the I/Es. For this carrier, overall technical and CRM performance ratings were based on a standard 4-point scale covering the full range of possible crew performance: unsatisfactory, satisfactory, standard, and above standard. The labels and precise meanings of each scale point were defined after several cycles of discussion between the I/Es and the research team. For example, satisfactory meant that

performance met or exceeded the FAA minimums, while standard meant that performance met or exceeded carrier performance standards, which were higher than FAA minimums.

### LOE Worksheets

Crew responses to normal, abnormal, and emergency situations within the LOE were assessed using structured worksheets for each event set in the scenario (for details see ATA, 1994; Hamman, Seamster, Smith, & Lofaro, 1991). These LOE worksheets simplified what could be a relatively complex evaluation process (see Figure 1) and provided instructors with a tool for making more reliable and valid ratings. The worksheets also helped evaluators standardize the assessment of LOE sessions and deliver a more balanced debrief to pilots that covered the CRM as well as the technical elements of each event set. Using worksheets based on each event set allowed the Instructor/Evaluator (I/E) to concentrate on a limited range of observable behaviors and on specific CRM and technical training objectives for each flight segment. However, the primary ways in which the worksheets helped facilitate the evaluation process were through structural changes and the addition of judgment anchors.

-----  
Insert Figure 1 here  
-----

LOE Worksheet Structural Changes. In response to I/E comments and suggestions, formatting for the LOE worksheets was simplified, streamlined, and augmented with specific judgment anchors for each evaluated skill during year three. The previous worksheet's dual-column format (Figure 2) was changed to a single-column format (Figure 3) that allowed the I/Es to progress linearly down the worksheet page. This revised format emphasized the correct order of the steps in the desired evaluation process. The previous mix of check-off boxes and numerical ratings was also changed to check boxes for all judgments. This emphasized the use of the standard rating process. The net effect of these changes was a cleaner, less cluttered evaluation

worksheet. The underlying rationale behind these changes was to clarify the rating process and reduce the I/Es level of workload during the evaluation event, thereby allowing additional time to make more reliable and valid assessments.

---

Insert Figures 2 and 3 here

---

The number of intermediate-level judgments was also increased in year 3 to encompass a more complete set of the specific tasks/skills listed in the carrier's AQP Program Audit Database (PADB). The item content of relevant tasks from the PADB was directly transferred to the evaluation form together with the corresponding PADB reference numbers. The reference numbers allowed the I/E to check the exact, formal definition of each evaluated skill. Since there were typically three to five skills evaluated during each event set during year 3, the set of evaluated skills included on the worksheet also may have been more complete than in previous years. For analysis, the technical ratings were averaged to obtain a composite judgment that was analyzed in the same manner as the technical judgments in years one and two.

Additional Judgment anchors. Also during this year, the I/Es requested additional guidance in how to evaluate *unsatisfactory*, *satisfactory*, *standard*, and *above standard* performance. Ultimately, a variation of a Behaviorally-Anchored Rating Scale (BARS) was developed that included concrete examples of each performance level for each specific skill (Smith & Kendall, 1963). These behavioral examples for each performance level were developed by a group of subject matter experts in the carrier's training department. Each performance level included a short description of the qualification standard for *standard* performance as well as brief examples of *unsatisfactory*, *satisfactory*, and *above standard* performance that were keyed to the specific skill or task being rated (Figure 4). For ease of reference, these performance levels were printed on the back of the preceding page of the LOE worksheet, so that they would be immediately available during the rating process.

-----  
Insert Figure 4 here  
-----

### I/E Rating Process

The I/Es were trained in a basic three-step evaluation process for using the worksheets to assess LOE performance. The evaluation process begins with the specific observable behaviors, moves to the intermediate judgments of tasks and skills, and combines this information to finally make the most general judgments of pilot and crew performance (Figure 1). More specifically, for each event set the I/Es were instructed to first evaluate the observable behaviors and rate them on a three-point scale of fully, partially, or not observed. The I/E then used these behaviors plus any other relevant information to evaluate the crew's technical and CRM skills using the standard four-point scale. The final step in the evaluation process was to use the ratings of technical and CRM skills to make overall evaluations of PIC, SIC, and crew performance for that event set using a 4-point scale.

The evaluation of structural validity focused on an examination of how well the I/Es were following this exact process for their LOE evaluations. Since LOEs were changed on a yearly cycle, the LOE evaluations were combined over each year for analysis. Since there were minor differences in the LOE content for each fleet due to differences in the aircraft and Standard Operating Procedures, the LOE data representing the rating process were initially analyzed separately for each fleet and then combined.

## RESULTS

### Reliability Data

IRR training sessions such as those described were conducted for each of the two fleets at a regional air carrier over a three-year period. Each index score is presented on a scale of 0.0 - 1.0

where scores approaching 1.0 are better. The sensitivity index was the latest index to be implemented and was not used until year 2. Therefore, it is not reported for year 1 (Table 1). Across all three years, congruency of the rating distributions and agreement across raters were generally acceptable (Tables 1, 2, and 3). Consistency was acceptable in year 2 but not in years 1 or 3. Sensitivity to small performance differences was disappointing in both year 2 and 3, and systematic differences continued to occur for some I/Es in each fleet.

-----  
Insert Tables 1, 2, and 3 here  
-----

Considering all the indexes, there is a general increase in scores from year 1 to year 2. From year 2 to year 3 (Tables 2 & 3), most of the indexes remained about the same but a few did decrease. However, it is difficult to compare across long time spans like the ones presented here for two reasons. The large turnover rate among the I/E population within our regional air carrier was magnified given that the I/E group was small to begin with (6-16 I/Es per fleet). Thus, one or two newly-minted I/Es grading significantly differently from the group could have drastically affected group performance.

Second, the content of the LOE that the I/Es were evaluating was normally quite different from year to year. The regional air carrier created a new LOE each year that was tailored to the most significant training or operational difficulties they were experiencing at that time. Therefore, the focal point of the items, topics, and situations was quite different from year to year.

Obviously it is important for any airline to be concerned about the reliability of their I/E population, however, the validity of the evaluation also requires attention. The validity of the evaluation process requires reliability among the I/Es but it also requires that the I/Es evaluate what they are supposed to be evaluating, in the manner the airline has trained them.

### Validity Data

Structural validity was used to estimate the validity of the evaluation process. The evaluations used specifically came from the LOE assessments of the pilots which were entered in the AQP performance database. These evaluations occurred annually for each pilot using the same scenario for all of them. There was an average of 120 LOEs conducted per year and 11 event sets per LOE. Each I/E typically evaluated multiple LOEs throughout the duration of the year. To assess the validity of the evaluation process, the technical and CRM ratings of the LOEs were regressed on the observable behavior ratings (for each event set) to determine the strength of connection between the observable behaviors and technical/CRM ratings. Second, the PIC, SIC, and crew evaluations were regressed on to the technical and CRM skills/task ratings. For each step in the path analysis, the percent of variance accounted for in each dependent variable was computed using a multiple regression analysis (i.e.  $R^2$ ) for the data from each fleet. The entries shown in Table 4 are the average of these  $R^2$  values for each step in the evaluation process across the two fleets.

---

Insert Table 4 here

---

In general, the connection between Observable Behaviors and technical and CRM evaluations is somewhat weaker than the connection between the technical CRM evaluations and pilot/crew evaluations. If the observable behaviors listed on the worksheet were a complete and fully diagnostic set of behaviors for each skill, we would expect the average validities in the first column to be much higher. In this case, low validities point to potentially serious problems in the development of the observable behaviors for event sets in the LOE or in the evaluation process.

However, each skill can be indexed by a wide set of possible behaviors. Space limitations on the worksheet and time limitations during the evaluation process limit the number of observable behaviors on the worksheet to a small set. This limitation may inherently limit the possible

validity values in the first column. The real issue may be how good these validity indexes can be given the constraints of the evaluation situation. Within these constraints, however, changes can be made which may increase the validity of the evaluations.

Although the observable behaviors on the evaluation form are carefully chosen to be the most important or diagnostic observable behaviors, I/E evaluations can be influenced by other observed behaviors of the crew. In contrast, the summary technical and CRM performance evaluations should be the focal point of evaluating both individual pilot and crew performance for that event set. Therefore, it is reasonable that these pilot and crew evaluations are more predictable from the evaluation of critical skills. These evaluations represent the relevant individual and crew skills that are assessed during each event set.

#### Changes in Validity after Changes in the Worksheet

The average structural validity of the ratings increased noticeably in year 3 (see Table 4). The average structural validity for year three was .48 compared to the average of .36 for year one and .33 for year two. Specifically, the event sets for the year 3 LOE displayed stronger relationships between the observable behaviors and the technical and CRM ratings than in years 1 and 2. The structural relationships between the technical and CRM ratings and the final PIC, SIC, and Crew evaluations also showed a modest overall improvement from previous years. These changes may have been due to a clearer format for the worksheet, more complete and specific judgment anchor definitions, or both.

### DISCUSSION

The results of this case study represent achievable evaluation standards that can be obtained with an IRR process in an operational pilot evaluation context with constrained resources of personnel and training time. Although a formal pre-post evaluation could not be done due to turnover in the I/E sample and changes in the materials from year to year, the results provide evidence of the reliability and validity of judgments among this group of evaluators.

#### Reliability

Having reliable I/E judgments means that these evaluations are stable or reproducible, both within and across raters. That is, to what degree can one I/E be counted on to give the same rating as another? Can we count on one I/E producing the same rating, given the same situation, at another time? Reliability of ratings across I/Es is critical for fair, unbiased evaluations. Since raters, like most experts, inherently wish to be stable in their evaluations over time, the across-rater component of reliability is usually the most problematic. Across-rater reliability should be amenable to appropriate training interventions such as the IRR training examined in this project.

Having an I/E who systematically rates pilots higher (i.e., a Santa Claus) or lower (i.e., an Axeman) than the rest of the I/E population is one source of unreliability across raters. The implication of having a “Santa Claus” providing ratings is that a pilot whose performance is actually below the airline’s minimum passing standards would get a passing evaluation and be allowed to fly the line when he or she should be receiving additional training. The cost of having an “Axeman” evaluating pilots is potentially less severe but costly nonetheless. An I/E grading too harshly can cost an airline money by providing additional training to pilots who actually meet the airline’s standards. Thus, either form of systematic differences among the evaluators is potentially costly. Additionally, if the I/Es are not providing reliable data about the performance of the pilot population, the airline cannot accurately analyze and detect pilot weaknesses or provide appropriate training interventions. Furthermore, if training interventions are implemented, they cannot be accurately assessed unless the evaluations made by the I/Es are reliable.

#### Benchmarks for Reliability

An important consideration in applying these metrics is the level of reliability shown by a particular cadre of I/Es and how that relates to standards of performance. For each of our metrics, data are now available on the range of achievable results within the constraints of a single day of training every six to eight months over a three-year period with a high rate of I/E turnover. Data from other groups that have different training and turnover constraints than the regional carrier

are also available. These data come from an industry workshop (Greenwood, Holt, and Boehm-Davis, 2000) and a major carrier that has implemented IRR training (Major carrier, personal communication, July 2000). Each group (the group of I/Es for each fleet at each carrier plus the group of workshop participants) is a different source of information on the achievable levels of these IRR indexes. The average value and the range of values found across these groups can be examined as a basis for setting minimum values or targets for the aviation industry.

Congruency. The values of congruency across all groups (the regional carrier, the major carrier, and the industry workshop) ranged from .67 to .86. The average across all groups for which we have congruency estimates is .76. Since the maximum value of this index is 1.0, these numbers are relatively high for the groups as a whole. In each group of raters, however, there were typically several I/Es who gave noticeably different distributions of ratings. The reasons for these rating distributions included an inadequate understanding of the rating scale or use of an idiosyncratic set of standards for making the ratings.

Systematic differences. Systematic differences among the raters were typically found, but the degree of this problem varied widely across the groups. Using a maximum value of 1.0 (reflecting *no* significant differences among the group), the range of values for systematic differences for the groups analyzed ranged from .40 to 1.0 (with an average across all groups of .75). Since systematic differences are found in most, if not all, groups of I/Es, the issue that must be addressed is the appropriate target value or goal for training. Should this goal be "zero tolerance" for any significant differences, or should some level of systematic differences among the I/Es be allowed? Clearly some groups were trained sufficiently well to achieve the goal of no raters showing significantly higher or lower ratings than the group. Given this evidence, an index value of 90% or higher (reflecting 10% or less of the group providing ratings that are significantly higher or lower than the group) is obviously achievable, but may require extensive training.

Consistency. The average level of consistency correlations varied widely among the groups, ranging from .16 to .80. The average across all groups for which we have consistency estimates

is .48. On a scale of 0 to 1, these values represent modest to strong inter-rater correlations in evaluations across items. Values of consistency in the .70s are clearly attainable, but the data from this project indicate that it may be difficult to achieve these values with the limited resources of a regional carrier since the baseline value is around .50.

Agreement. For agreement, the range of values for all groups was found to be .58 to .88. The average across all groups for agreement estimates is .76. On a scale from 0 to 1, these values indicate medium to high agreement across all items. However, agreement for specific items is often inadequate and is a useful focus of group discussion in the IRR training. At the regional carrier, we used a value of .70 to target individual items for group discussion and resolution. Clearly values of .70-.80 or above for average item-level agreement can be achieved in an operational context, even within the constraints of a regional carrier.

Sensitivity. The average level of sensitivity estimates also varied widely among the groups, ranging from -.04 to .20. The average across all groups for which we have sensitivity estimates is .07. At the regional carrier, levels of sensitivity were quite low across the three years, but higher values were found for participants in the industry workshop (Greenwood, Holt, and Boehm-Davis, 2000; Williams, Holt, & Boehm-Davis, 1997). There are at least two distinct possible causes for very low sensitivity estimates.

First, the initial specification of the different performance levels by SMEs can be a methodological problem in getting accurate sensitivity estimates for a training videotape. If these specifications are inaccurate, the sensitivity estimate from the IRR analysis could be biased below its true value. Since establishing sensitivity requires accurately specifying levels of performance for segments of the training videotape, the differences in sensitivity found among groups so far may be due to having a better or worse specification of performance levels. The performance levels for the tape used for the industry workshop had been reviewed by several SMEs (Greenwood, Holt, & Boehm-Davis, 2000; Williams, Holt, & Boehm-Davis, 1997); sensitivity

was found to .36. Clearly a level of .36 can be achieved even for a diverse I/E population in the context of a 2-day workshop.

However, in the work at the regional carrier, few SMEs, sometimes only one, was available for determining performance levels prior to the training event. Relying on the judgment of a single or small number of SMEs is particularly apt to be a poor choice because of the tendency of the SME/s to think his or her evaluations are universal. This "false consensus effect" can only be corrected by having multiple SMEs evaluate performance. This is a similar problem to establishing a good gold standard for evaluating performance. Baker & Dismukes (in press) discuss establishing a gold standard of real performance levels for videotaped flight segments to be used for training evaluators. The issue of accurately determining performance levels is also important for IRR because inaccurate specifications of performance can limit sensitivity values and this may be one reason the sensitivity values for the regional carrier in this project were typically low.

A second possibility is that the performance levels determined by the SME(s) were correct but that the distinction between closely related performance levels (such as a "2" and "3" level performance) is inherently very difficult to judge. Although distinguishing extremely good from extremely bad performance is quite easy even for naïve evaluators, reliably distinguishing closely related performance levels may be much more difficult. Ensuring the precision of such judgments may require careful attention to judgment criteria or anchors and the judgment process for the each specific type of rating. This effort should be targeted at the discrimination that is considered most important for pilot training and evaluation. For the carrier in this case study, the distinction between "standard" performance and "satisfactory" performance was important because the "satisfactory" performance had to be debriefed or retrained to become carrier standard. Wherever the critical distinction lies on the evaluation scale, the evaluation tools as well as the evaluator training should be carefully examined to ensure sensitivity in evaluations.

Interpretation of the Proposed Benchmarks. These tentative baselines should be viewed cautiously as they are taken from very different samples. The regional carrier in this project experienced high I/E turnover during this period, which would adversely affect almost all aspects of reliability. The IRR scores may also depend to some extent on the particular group of I/Es who are analyzed in each session. Different airlines may recruit different populations of I/Es or have different methods and standards for training their I/Es; they may therefore find different initial values for these indexes. Under more optimal conditions of low I/E turnover and more total time available for relevant training, higher levels of these IRR indexes should be achievable. Comparable data from a major carrier (Major carrier, personal communication, July 2000) indicated a value of .92 for systematic differences, .76 for congruity, .73 for consistency, and .87 for agreement (sensitivity was not evaluated in that training). If carriers are willing to share results of IRR training with other carriers, the conditions and training methods that produce higher levels of IRR could be determined.

As more airlines implement such training, it should become easier to estimate some industry wide benchmarks for each index that would be attainable and necessary for successful and reliable I/E evaluations (see Williams et al, 1997 for more discussion of proposed benchmarks). Clearly there may be a tradeoff or a point of diminishing returns between resources invested in I/E training and the resulting levels of reliability. The contrasting data from the regional carrier, the workshop, and a major domestic carrier illustrate this point.

The reliability data at the regional carrier suggest that despite high rates of turnover in the I/E cadre and limited training time, reasonably high levels of agreement and congruency could be achieved with careful attention to the details of the evaluation process. However, the levels of inter-rater consistency were only moderate and sensitivity was poor. Increasing consistency may require more detailed item-by-item feedback on the reasons and standards for each assessment. Establishing a common frame of reference for judgment should help generate better consistency by making the higher or lower evaluations of each item more comparable across I/Es. Better

evaluations of sensitivity may require allocation of multiple SMEs to review and evaluate segments of the training tape similar to the process of establishing gold standards for performance discussed by Baker & Dismukes (in press).

### Validity

The structural validity data for year three compared to years one and two suggested increased validity through the use of refined worksheets, better judgment anchors, and extensive feedback based on qualitative results. This result implies that improving rater evaluations should focus on the human factors of the evaluation instrument and evaluation process as well as on rater training and calibration.

### Importance of Structural Validity

Structural validity is important to assure that the I/Es are evaluating the appropriate behavior in the prescribed manner and it also can aid the reliability of the assessments. The improvement of the structural validity in year 3 displays the effectiveness of the worksheet modifications that were made. Both of the modifications, change in structure and the addition of the judgment anchors, are changes that other airlines and organizations can do to enhance the validity of their evaluators' assessments.

Importance of LOE Worksheet Structure Design. The LOE worksheet is the primary evaluation tool and must be designed to facilitate the evaluation process. First, the evaluation flow must be clearly specified on the sheet in a simple and direct manner. A linear top-to-bottom organization of ratings in chronological order for each event set worked well. Second, the set of evaluations must be made as simple as possible. In particular, the workload of the raters during the evaluation event must be considered when designing the response format. Here, a consistent response format emphasizing the standard four-point scale worked well. Any auxiliary ratings must also be made as simple as possible. In Figure 3, for example, the auxiliary judgments of reason codes for poor performance were made by simple check off of letters representing each code.

Importance of Judgment Anchors. Giving explicit and concrete examples of each level of a judgment scale is one method to improve ratings. The I/Es requested these examples and providing them did seem to facilitate the structural validity of the ratings. This extends general research on the use of examples and judgment anchors for stabilizing ratings to this domain. For this project, the examples were constructed and added to the reverse side of the worksheets for ease of access during the high-workload evaluation session. If evaluations are computer-based, pop-up examples could be implemented in an interactive fashion during the evaluation session. However, access to the examples serving as judgment anchors must be quick and easy or raters may not use them. The examples must be available during the judgment process in the evaluation setting, and the process of accessing the examples must not contribute to evaluator workload in a high workload environment.

Importance of IRR Post-Session Feedback Form. The changes in the worksheets and more extensive judgment anchors implemented in year three of this study were heavily influenced by past comments and data recorded from past IRR training sessions. Thus, the post-session feedback form proved to be an important tool. Specifically, it functioned as an effective way to document the judgment standards and criteria that evolved from the IRR sessions. It also proved to be an effective method of following up on a wide variety of unresolved and important issues for the regional airline. The form summarized the qualitative and quantitative results brought out in the discussion during the IRR training session. In part, this feedback led to a broad set of suggested improvements in the definition, training, and evaluation of pilot performance. These issues were brought to the attention of the training department, flight safety, simulator technicians, and company policy makers, as appropriate. For example, company flight standards for dividing Pilot Flying and Pilot Not Flying duties were clarified to require both a clear briefing of the intended division of duties plus a clear enactment of each set of duties by each pilot.

#### Future Directions

The results of this study are encouraging but not definitive. Further research must evaluate the effectiveness of the IRR training by gathering and analyzing either pre-post training evaluations or by comparing the evaluation results of a trained group of I/Es versus an untrained group. Having trained and untrained groups of I/Es was not feasible in this study because only two groups of I/Es existed at the airline (one for each fleet) and the evaluation of Advanced CRM required that the I/Es in both fleets be trained.

Evaluating the effects of training by pre- and post-training assessments is an alternative approach. In evaluating training effectiveness, the initial I/E groups should be large enough to provide an adequate sample size when the predictable effects of attrition are taken into account. However, one initial effect of being shown evidence of poor calibration may be a re-examination of the I/E judgment framework. This self-criticism may lead to a short-term variation in judgments while the old framework is abandoned or modified and a new framework for judgment is being consolidated. Therefore, it is important that the post-training assessments include long-term follow-up. This long-term follow-up evidence should include not only future evaluations of test segments but also the distribution of judgments made during normal on-the-job evaluations of crews.

Future work should also address the levels of IRR benchmarks that can be achieved with specific I/E populations and training methods. The accumulation of this evidence will help set practical and achievable standards for crew evaluation in the aviation community. The challenge is to find ways to effectively communicate these findings and new ideas to all segments of the aviation community without violating the confidentiality of individual I/E results or proprietary company information.

#### ACKNOWLEDGEMENTS

This research was supported by a grant from the Federal Aviation Administration, AAR-100, through Grant Number 94-G-034. The views expressed in this paper represent the views of the authors and not that of the federal government. The authors thank Thomas L. Seamster for his

work in developing the proceduralized training program. The authors also express their gratitude to Captain Kim Schulz of Atlantic Coast Airlines and Captain William R. Hamman of United Air Lines for their assistance in developing and coordinating the implementation of this research program.

## REFERENCES

- ATA (1994). *Line Operational Simulations: LOFT Scenario Design, Conduct and Validation*. LOFT Design Focus Group, AQP Subcommittee Report, November 2, 1994.
- Baker, D.P. & Dismukes, R.K. (in press) Training raters to assess crew performance: theory practice and future directions. *International Journal of Aviation Psychology*.
- Baker, D.P. & Mulqueen, C. (1999). Pilot instructor/evaluator rater training: guidelines for development. In Proceedings of the Tenth international symposium on aviation psychology. Columbus, OH: Ohio State University.
- Bernardin, H.J., & Buckley, M.R. (1981). Strategies in rater training. *Academy of management review*, 6, 205-212.
- Birnback, R.A., & Longridge, T.M. (1993). The regulatory perspective. In E.L. Wiener, B.G. Kanki, & R.L. Helmreich (Eds.), *Cockpit resource management*, New York: Academic Press.
- Boehm-Davis, D. A., Holt, R. W., & Seamster, T. (in press). Resource management in aviation: Two airlines' experience. To appear in E. Salas, C. A. Bowers, and E. Edens (Eds.), *Applying Resource Management in Organizations: A Guide for Training Professionals*, NJ: Lawrence Erlbaum Associates.
- Borman, W.C. (1975). Effects of instruction to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of applied psychology*, 60, 556-560.
- Cohen, J. & Cohen, P. (1983) *Applied multiple regression/correlation analysis for the behavioral sciences* (2<sup>nd</sup> ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues*. Houghton Mifflin Co.
- George Mason University (1996). Developing and evaluating CRM procedures for a regional air carrier phase I report. Federal Aviation Administration, Office of the Chief Scientific and Technical Advisor for Human Factors, Washington, DC.
- Greenwood, D. M., Holt, R. W., and Boehm-Davis, D. A. (2000). Training instructor pilots to evaluate air crew performance in a workshop setting, *Technical Report*, Fairfax, VA: George Mason University.
- Hamman, W.R., Seamster, T.L., Smith, K.M., & Lofaro, R.J. (1991). The future of LOFT scenario design and validation. *Proceedings of the 6<sup>th</sup> International Symposium on Aviation Psychology*, 589-594.

Hays, W. L. (1981) *Statistics*. New York: Holt, Rinehart and Winston.

Holt, R. W. (in press). *Scientific Information Systems*. Aldershot: Ashgate.

Holt, R.W., Meiman, E., & Seamster, T.L. (1996). Evaluation of aircraft pilot team performance. *Proceeding of the human factors and ergonomics society 40<sup>th</sup> annual meeting*. Philadelphia, PA.

Jako, R.A. & Murphy, K.R. (1991). Distributional ratings, judgment decomposition, and their impact on interrater agreement and rating accuracy. *Journal of applied psychology*, 75, 500-505.

James, L.R., Demaree, R.G., Wolfe, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of applied psychology*, 69, 85-98.

James, L.R., Demaree, R.G., Wolfe, G. (1993).  $r_{wg}$ : An assessment of within-group interrater agreement. *Journal of applied psychology*, 78, 306-309.

Johnson, P. J. & Goldsmith, T. E. (1998). The importance of quality data in evaluating aircrew performance. FAA Technical Report.

Law, J.R. & Sherman, P.J. (1995). Do raters agree? Assessing inter-rater agreement in the evaluation of air crew resource management skills. In Proceedings of the Eighth International Symposium on Aviation Psychology. Columbus, OH: Ohio State University.

McGraw, Kenneth O; Wong, S. P. Forming inferences about some intraclass correlations coefficients: Correction. *Psychological Methods*. Vol.1(4), Dec 1996, 390.

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

Pedhazur, E.J.; Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach* (student ed.). Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc. xiv, 819.

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*. Vol 3(1), 69-82.

Saal, Frank E; Downey, Ronald G; Lahey, Mary A. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*. Vol 88(2), Sep 1980, 413-428.

Smith, P. C. & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47(2) 149-155.

Viswesvaran, C. Ones, D.S., & Schmidt, F.L. (1997). Comparative analysis of the reliability of job performance ratings. *Journal of applied psychology*, 81, 557-574.

Williams, D. M., Holt, R. W., and Boehm-Davis, D. A. (1997) *Training for inter-rater reliability: Baselines and benchmarks*. In R. S. Jensen & L. Rakovan (Eds.), *Proceedings of the*

*Ninth International Symposium on Aviation Psychology*, (pp. 514-519). Columbus, OH: The Ohio State University.

Woehr, D.J. & Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of occupational and organizational psychology*, 67, 189-205.

Table 1. IRR Training Results across Two Different I/E Groups for Year 1.

<b>Year 1</b>					
<b>Fleet 1 Ratings (N=6)</b>	<b>Systematic Differences</b>	<b>Congruency</b>	<b>Consistency</b>	<b>Sensitivity</b>	<b>Agreement</b>
3-point	1.0	.71	.56	NA	.67
4-point	.40	.67	.36	NA	.88
<b>Fleet 2 (N=8)</b>					
3-point	.71	.69	.43	NA	.58
4-point	.71	.72	.46	NA	.86

Table 2. IRR Training Results across Two Different I/E Groups for Year 2.

<b>Year 2</b>					
<b>Fleet 1 Ratings (N=14)</b>	<b>Systematic Differences</b>	<b>Congruency</b>	<b>Consistency</b>	<b>Sensitivity</b>	<b>Agreement</b>
3-point	.85	.86	.80	.13	.76
4-point	.54	.76	.75	.01	.85
<b>Fleet 2 (N=12)</b>					
3-point	.83	.78	.67	.09	.65
4-point	.83	.81	.68	.01	.84

Table 3. IRR Training Results across Two Different I/E Groups for Year 3.

Year 3					
Fleet 1 Ratings (N=16)	Systematic Differences	Congruency	Consistency	Sensitivity	Agreement
3-point	.79	.77	.36	-.04	.64
4-point	.71	.73	.16	.09	.82
Fleet 2 (N=14)					
3-point	.93	.82	.23	.05	.71
4-point	.69	.78	.28	.20	.84

Table 4. Structural Validity Results ( $R^2$ ) for Three Years of Evaluations.

Year:	Average Structural Validity		
	Predicting technical & CRM Ratings from Observable Behaviors	Predicting PIC, SIC & Crew Ratings from Technical and CRM Ratings	Overall Average Percent of Variance
1	.21	.51	.36
2	.21	.46	.33
3	.37	.58	.48

### Figure Captions

Figure 1. Standard LOE worksheet evaluation process of items beginning at the most specific, lowest level of evaluations (OBs) and finishing with the most general, highest level of evaluations (crew rating).

Figure 2. Formatting of a portion of one event set of the original worksheet (used in years 1 and 2).

Figure 3. Formatting of a portion of one event set of the revised worksheet (used in year 3).

Figure 4. The judgment anchors for the "cockpit preparation" task shown in the revised worksheet example (used in year 3).

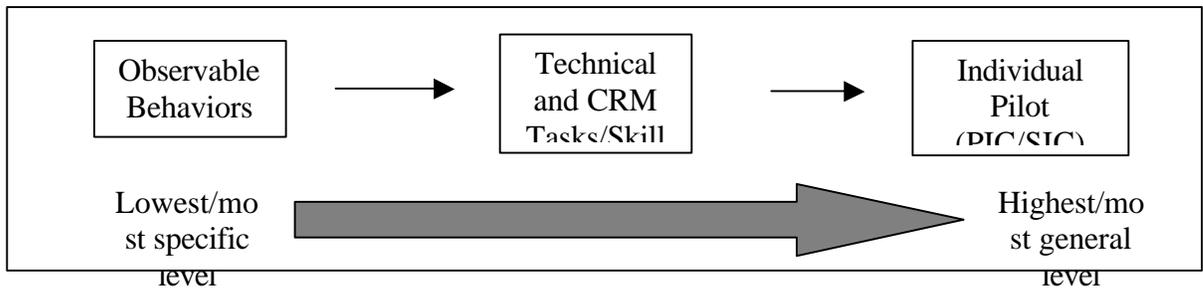


Figure 1

<b>LEG 1 EVENT SET 1 (Pre Departure through Taxi)</b>							
<b>CRM OBSERVABLE BEHAVIORS</b> CHECK ✓ ONLY ONE BOX FOR EACH BEHAVIOR BELOW	FULLY OBSERVED	PARTIALLY OBSERVED	NOT OBSERVED	<b>TECHNICAL AND CRM RATINGS</b> <b>RATE FROM 1 to 4</b> <b>EACH TECHNICAL AND CRM ITEM</b>			<b>RATE 1-4</b>
<b>TEAM MANAGEMENT OBSERVABLE BEHAVIORS</b>				<b>TECH:</b> Interpretation of Airport Analysis			
Crew performs complete briefing to include summer operations SOP				<b>TECH:</b> Handling of abnormal start			
Crew discusses reasons for hot start and briefs plan for restart				<b>CRM:</b> Crew briefing			
<b>BRIEFING OBSERVABLE BEHAVIORS</b>				<b>OVERALL Event Set 1.1</b> <b>RATE PIC, SIC &amp; CREW 1-4</b> <small>(see above)</small>			PIC SIC CREW
Crew discusses the need to communicate and keep each other in the loop				<b>REASON CODE</b> <small>(see above)</small> <b>FOR ANY OVERALL RATING NOT A 3</b>			PIC SIC CREW
Crew discusses clearance brief items				<b>REPEATS REQUIRED?</b>			PIC SIC CREW
				<b>CHECK ✓ BOX OF PILOT FLYING</b>			PIC SIC

Figure 2

<b>LEG 1 EVENT SET 1 (Pre-Departure to Taxi)</b>														
<b>OBSERVABLE BEHAVIORS</b> ✓ CHECK ✓ ONLY ONE BOX FOR EACH BEHAVIOR BELOW						NOT OBSERVED	PARTIALLY OBSERVED	FULLY OBSERVED	MISSED Observation					
Crew discusses need for takeoff alternate.														
Crew discusses low visibility takeoff procedure.														
Crew briefs F/A about turbulence prior to T.O.														
<b>TASKS</b> ✓ CHECK ✓ ONLY ONE BOX FOR EACH TASK BELOW						Repeat/ Unsat. 1	Debrief 2	Standard 3	Above Standard 4					
2.1	Cockpit preparation: checklists, W&B, performance.													
2.1.8. 3	Clearance briefing includes the relevant items.													
2.1.9	Dispatch Release properly amended due to significant reroute.													
<b>OVERALL EVENT SET RATINGS</b> ✓ Check ✓ only <u>ONE</u> box for the Overall ratings below <u>and</u> ✓ Check ✓ up to 3 Reason Codes						Repeat/ Unsat. 1	Debrief 2	Standard 3	Above Standard 4					
<b>PIC Overall</b>														
PIC Reason Codes for any overall rating NOT a 3 ( <i>see above codes</i> )						A	J	K	P	T	C	S	D	W
<b>SIC Overall</b>														
SIC Reason Codes for any overall rating NOT a 3 ( <i>see above codes</i> )						A	J	K	P	T	C	S	D	W
<b>CRM Overall</b>														
<b>CREW Overall</b>														
CREW Reason Codes for any overall rating NOT a 3 ( <i>see above codes</i> )						A	J	K	P	T	C	S	D	W

Figure 3

Task #	Task Description	Repeat / Unsat. 1	Debrief 2	Standard 3	Above Standard 4
2.1	Cockpit Preparation, Checklists, W&B, Performance	Checklists not completed or with omissions, or completed by memory, W&B and Performance computations are not completed, or completed with gross errors	All checklists completed without omission, but using non standard phraseology ,and not completed by memory, W&B and Performance computations are completed with minor errors	All checklists completed without omission, and not completed by memory, W&B and Performance computations are completed without errors	One crewmember catches errors or omissions, and helps the other crewmember through the procedure.

Figure 4