AMERICAN INSTITUTES FOR RESEARCH

CRM ASSESSMENT: DETERMINING THE GENERALIZATION OF RATER CALIBRATION TRAINING

SUMMARY OF RESEARCH REPORT: GOLD STANDARDS TRAINING

Principal Investigator: David P. Baker, Ph.D.

Period of Performance: February 1998 – June 2002

Prepared by:

David P. Baker, Ph.D.
American Institutes for Research
1000 Thomas Jefferson Street, N.W.
Washington, DC 20007

Prepared for:

National Aeronautics and Space Administration
Ames Research Center
Moffett Field, CA 94035-1000

Grant Number:

NCC 2-1084

AMERICAN INSTITUTES FOR RESEARCH

09 September 2002

NASA Center for Aerospace Information (CASI)
Document Processing Section
7121 Standard Drive
Hanover, MD 21076

Dear CASI,

Enclosed is one original copy of a Summary of Research report entitled, "Gold Standards Training." This document is submitted in fulfillment of the final reporting requirement for NASA Ames Cooperative Agreement Number NCC 2-1084. The enclosed document summarizes the research conducted under this agreement, which studied the effectiveness of different strategies for training pilot instructors to assess crew performance. Under the Federal Aviation Administration's (FAA) Advanced Qualification Program (AQP), pilot instructor rater training is required, and instructors must be calibrated periodically to ensure the reliability and validity of crew performance assessments.

To address this critical need, the American Institutes for Research (AIR) reviewed the relevant research on rater training and, based on "best practices" from this research, developed a new strategy for training pilot instructors to assess crew performance. In addition, we explored new statistical techniques for assessing the effectiveness of pilot instructor rater training. Results of our research were described in numerous publications and conference presentations that are summarized in the enclosed Summary of Research report.

Questions regarding technical activities conducted under this cooperative agreement may be addressed to me, the Principal Investigator. Business related questions may be addressed to Mr. Thomas Jesulaitis, AIR Contract and Grants Officer. I can be reached at (202) 342-5036 or via e-mail at dbaker@air.org. Tom can be reached at (202) 342-5031 or via e-mail at tjesulaitis@air.org. We appreciate this opportunity to support NASA Ames in this important area of research.

Sincerely,

David P. Baker, Ph.D.

CC: TO
AGO
GO
CTO

CRM ASSESSMENT: DETERMINING THE GENERALIZATION OF RATER CALIBRATION TRAINING

SUMMARY OF RESEARCH REPORT: GOLD STANDARDS TRAINING

Principal Investigator: David P. Baker, Ph.D.

Period of Performance: February 1998 – June 2002

Prepared by:

David P. Baker, Ph.D.
American Institutes for Research
1000 Thomas Jefferson Street, N.W.
Washington, DC 20007

Prepared for:

National Aeronautics and Space Administration
Ames Research Center
Moffett Field, CA 94035-1000

Grant Number:

NCC 2-1084

TABLE OF CONTENTS

SUMMARY OF RESEARCH REPORT

The extent to which pilot instructors are trained to assess crew resource management (CRM) skills accurately during Line-Oriented Flight Training (LOFT) and Line Operational Evaluation (LOE) scenarios is critical. Pilot instructors must make accurate performance ratings to ensure that proper feedback is provided to flight crews and appropriate decisions are made regarding certification to fly the line. Furthermore, the Federal Aviation Administration's (FAA) Advanced Qualification Program (AQP) requires that instructors be trained explicitly to evaluate both technical and CRM performance (i.e., rater training) and also requires that proficiency and standardization of instructors be verified periodically.

To address the critical need for effective pilot instructor training, the American Institutes for Research (AIR) reviewed the relevant research on rater training and, based on "best practices" from this research, developed a new strategy for training pilot instructors to assess crew performance. In addition, we explored new statistical techniques for assessing the effectiveness of pilot instructor training. The results of our research are briefly summarized below. This summary is followed by abstracts of articles and book chapters published under this grant.

Rater Training Research

A review of relevant research found that there are four strategies for training raters to make accurate and reliable performance assessments (Baker, Mulqueen, & Dismukes, 2001). These training approaches are: Rater Error Training (RET), Performance Dimension Training (PDT), Behavioral Observation Training (BOT), and Frame-Of-Reference (FOR) training. With the exception of BOT, each has been widely studied. A brief overview of the goals and methods of each strategy is presented in Table 1 below.

Regarding the effectiveness of the four rater-training strategies mentioned above, the research indicated that Frame-Of-Reference (FOR) training was the most effective single strategy for training raters to make accurate ratings. Behavioral Observation Training (BOT) also was found to have a moderate to large effect on rating accuracy. However, caution is warranted regarding findings for BOT, since there are very few studies that have investigated the effectiveness of this training strategy to date.

In addition to examining the effectiveness of each rater-training strategy, the literature review also examined the effectiveness of combinations of the different rater-training strategies. This literature indicated that combinations of strategies that are individually effective lead to even higher gains in accuracy. Finally, the literature suggested that a combination of group discussion with significant opportunities for practice and feedback represent the most effective training methods (Smith, 1984).

Table 1. Rater training strategies.

Strategy	Goals	Method
Rater Error Training (RET)	Reduce rating errors; produce more normal distributed ratings.	Familiarize raters with common rating errors (e.g., halo, leniency).
Performance Dimension Training (PDT)	Increase rating accuracy by facilitating dimension-relevant evaluations.	Familiarize raters with performance dimension and rating scales.
Behavioral Observation Training (BOT)	Increase rating accuracy by focusing on the observation of behavior.	Utilize strategies that focus on observing and recording behavior (e.g., note-taking).
Frame-of-Reference Training (FOR)	Increase rating accuracy by focusing on the different levels of performance.	Provide raters with different standards of performance on dimensions. Include rating practice and feedback.

Based on our literature review, we presented what we believe, is the most effective approach for training instructors to assess aircrew performance during LOFTs and LOEs: Gold Standards Training. Gold standards training combines tested and proven strategies from the methods described in Table 1.

Gold Standards Training

In this section, we describe what a combination of the best practices from different rater training strategies might look like. We refer to this training as "gold standards" training, because it combines the most desirable characteristics of BOT and FOR, and relies upon gold standards for providing pilot instructors feedback about their rating accuracy. In Appendix A, we present a course design guideline for developing gold standards training. Instructional design experts can use this document to develop this training.

Scenario Review

First a detailed review of the LOE or LOFT scenario(s) to be evaluated should be conducted. In addition to the scenario, the review should cover each of the event sets that comprise the scenario and the CRM and technical skills to be evaluated. In cases where pilot instructors are being trained for the first time, this review should also include a detailed explanation of any grade sheets used to assess CRM and technical performance. Review of the various types ratings to be made (e.g., CRM, technical, event set, etc.) and any grading rules that apply (e.g., cases where certain behavioral observations lead to specific CRM ratings).

Performance Standards Review

The performance standards for each technical and CRM skill to be assessed should be reviewed. This review is the first step in developing consistent standards across new pilot instructors for evaluating aircrew performance during scenario-based training. Information regarding the requirements for successful crew performance on each scenario event set is often found in or can be developed from the scripts that describe the LOE or LOFT. Information from these scripts can be leveraged to develop specific examples of different performance levels on the grade sheet.

Observation skills training

To ensure pilot instructors accurately observe technical and CRM behaviors during LOE or LOFT, gold standards training should include behavioral observation training. BOT is based on the premise that there is a significant difference between the processes involved in observation and the processes involved in evaluation (Thornton & Zorich, 1980). According to this view, observation processes encompass the detection, perception, and recall of behavioral events, while evaluation processes include categorizing, integrating, and evaluating information.

Observation training should include both a discussion and a practice and feedback component. First, discussion should focus on the nature of a good observation (i.e., specific, behavioral, verifiable) and how to accurately observe a team's performance during a scenario. Second, observation training should include opportunities for practice and feedback. The research on rater training suggests that practice and feedback is critical for training transfer (Smith, 1986). Therefore, instructors should be shown videotapes of teams performing the scenario for the purpose of practicing their observational skills. These videotapes should be annotated with detailed observations from experts about the specific behaviors exhibited by the aircrews shown on the videos and how those behaviors are best interpreted. This annotation provides detailed feedback to the instructors so they can compare what they observed or failed to observe and how they interpreted their observations to observations and interpretations of experts.

Rating practice

A key component of gold standards training is practicing the rating task. Ideally, this practice involves rating the videotaped performance of aircrews performing events from the LOE or LOFT scenario(s) that will be rated by pilot instructors in the future. Practice videos should display a range of aircrew performance. Here, we recommend including a minimum of at least three practice videotapes displaying excellent, average, and poor technical and CRM skills on each event set to be rated.

Table 2. Gold standard example.

SCENARIO EVENT SET 3

TRIGGER: System malfunction during climb-out. The malfunction is the Leading Edge (LE) Slat fails to retract in icing conditions.

EVENT SET GRADES	GOLD STANDARD RATINGS	GOLD STANDARD RATIONALES
Teamwork	3	Teamwork behaviors observed: The crew requested time on the runway for engine run-up. The captain watched outside the aircraft for sliding during engine run-up while the first officer set throttles to 70%. The first officer verbalized a plan for handling the LE Slat problem. The captain suggested that the crew wait to deal with the LE Slat problem until the aircraft was on its assigned heading. The captain handled the LE Slat Transit Light – On checklist while the first officer flew and talked to air traffic control.

Analyze rating data

Course instructors perform this task. Here, instructors analyze the practice ratings and prepare materials for providing feedback to new pilot instructors. At that heart of these analyses is the comparison of the new pilot instructor ratings for the videotapes to "gold standards" that have been developed for each practice video. Essentially, gold standards are "true scores" that have been developed by expert pilot instructors for each videotape used in training. The primary focus of gold standards training is to teach new pilot instructors to rate teamwork skills more like expert instructors. Research indicates that such training increases accuracy. Specific methodologies for developing gold standards have been presented in the literature (Baker, Swezey, & Dismukes, 1998). An aviation example of gold standards appears in Table 2.

Regarding the actual data analysis, Goldsmith and Johnson (in press) provide an informative discussion of the application of statistical methods for analyzing trainee data using gold standards. Specially, they describe measures of referent reliability and instructor accuracy and provide formulas for calculating these methods. Holt and his colleagues (Holt, Hansberger, Boehm-Davis, in press) have also developed an automated tool for conducting such analyses.

Performance feedback

Gold standards training should include feedback based upon the results of the analyses. In addition to data-driven feedback, qualitative feedback should be provided using the expert rationales that are developed for each gold standard. The research evidence demonstrates the importance of using gold standards for training instructors to assess technical and CRM skills in the same way as instructor experts (Bernardin & Buckley, 1981). Furthermore, when multiple cadres of instructors require training, gold standards ensure that consist feedback is provided across instructor classes. As a result, greater reliability and accuracy should be observed not only within each instructor class but also across classes (Baker & Dismukes, in press).

Statistical Techniques for Assessing Rater Training Effectiveness

In addition to the development of gold standard training, we investigated the utility of different statistical techniques for assessing pilot instructor training effectiveness. Specifically, Mulqueen and his colleagues (Mulqueen & Baker, 1999; Mulqueen, Baker, & Dismukes, in press) explored the benefits to using multifacet Rasch analysis to assess pilot instructor training programs.

Multifacet Rasch analysis, which is derived from item response theory, allows researchers to assess the effects of multiple factors on pilot instructor accuracy. (Traditional analyses do not have this ability.) These factors include crew abilities, individual rater tendencies, scenario difficulty, and the difficulty of evaluating particular CRM skills. For example, Rasch analysis allows researchers to assess the quality of specific flight scenarios used during training. It can also indicate which skills pilot instructors have the most (and least) trouble evaluating. Moreover, this analysis allows researchers to examine the combined effects of these factors on pilot instructor accuracy. This allows trainers to determine, for example, if certain raters are having difficulty assessing particular CRM skills, or if a pilot instructor was too harsh or lenient in rating a particular flight crew.

It should be noted that there are several current limitations associated with Rasch analysis. First, producing the analysis is cumbersome and difficult to learn. This makes its use during a pilot instructor rater-training class difficult. Second, item response theory, on which Rasch analysis is based, is not well known or understood among researchers and training participants. In addition, the feedback from the analysis may be difficult for pilot instructors to understand. These factors represent drawbacks to using Rasch analysis, particularly to examine pilot instructor ratings during a training class.

Despite these limitations, the utility of using Rasch analysis in the assessment of training effectiveness is two-fold. First, it can provide specific information about the accuracy of individual raters, which allows trainers to tailor their feedback to particular pilot instructors. Second, the information provided about scenario difficulty and skill evaluation difficulty can be especially useful when developing new scenarios or rating forms. Such information can indicate if an existing scenario is too easy or too difficult and needs to be modified. Rasch analysis is useful both in the training process itself and as a tool for developing new training materials.

Summary Comments

Though this grant, AIR has been able to advance the science of training pilot instructors to assess crew performance. In addition, we have also explored new ways to analyze data collected from a training class. We believe that both projects represent significant achievements in the area of pilot instructor training.

Rasch analysis represents a new and more complete way to evaluate pilot instructor training. Despite its limitations, this approach provides more detailed information about multiple aspects of a training program than traditional analyses. The amount of information obtained using this analysis, especially with regard to individual rater tendencies and scenario difficultly, makes it a promising avenue for future research.

AIR's gold standard training incorporates the "best practices" of several well-researched training strategies. The aspects of the training strategies incorporated into gold standards training have been empirically shown to improve rater accuracy. In addition, there is indirect evidence to show that gold standard training should eliminate the problem of inconsistent rating norms developing between training classes. Gold standard training should result in consistent improvements in pilot instructor accuracy across training classes, not just within individual classes. Based on these findings, we recommend that gold standards training become the method for training pilot instructors to assess crew performance within the airline industry.

Grant Abstracts

Gold Standards Training

David P. Baker

American Institutes for Research

R. Key Dismukes

NASA Ames Research Center

Background and Applications

Training teamwork skills is increasingly important in a wide variety of organizations. For example, within commercial aviation, effective teamwork is critical on the flight deck. In this industry, where the consequences of error are extreme, the vast majority of incidents and accidents have been attributed to breakdowns in the teamwork of aircrew members (Helmreich, Weiner, & Kanki, 1993). As a result, commercial aviation has been a leading contributor to the development of effective team training, or crew resource management (CRM) as it is known in the airline industry.

An important feature of most team training programs is their strong reliance upon scenario-based training for skills practice. Essentially, scenarios are job simulations in which identifiable events are embedded to elicit specific team behaviors (Smith-Jentsch, Johnston, & Payne, 1998). An instructor(s) observes the team's performance during the simulation and rates the team on specific teamwork skills (Goldsmith and Johnson, in press).

A critical factor in scenario-based training is the instructor. Inevitably, the effectiveness of scenario-based training rests upon the ability of the instructor to observe relevant behavior and make an accurate evaluation of a team's teamwork skills. As Birnbach and Longridge (1993) noted, scenario-based training can only be effective if instructor ratings are accurate and reliable. The most direct and efficient method for ensuring that instructors will achieve these objectives is to provide them with rater training. Formal rater training can serve to familiarize instructors with the scenario, the scenario events, and the skills to be assessed.

This chapter presents, what we believe, is the most effective approach for training instructors to assess teamwork during scenario-based training: Gold Standards Training. Gold standards training combines tested and proven rater training strategies from the domain of performance appraisal – Behavioral Observation Training and Frame-of-Reference Training – to produce a methodology for effectively training instructors to evaluate team performance.

Reference

Baker, D. P., & Dismukes, R. K. (submitted for publication). Gold standards training. *Handbook on Human Factors and Ergonomic Methods*.

A Framework for Understanding Crew

Performance Assessment Issues

David P. Baker

American Institutes for Research

R. Key Dismukes

NASA Ames Research Center

Abstract

The focus of this special issue is on training pilot instructors to assess crew performance. In this opening article we attempt to set the stage for the other articles in this volume by introducing a framework for understanding crew performance assessment. We use this framework to outline issues that should be addressed when training pilot instructors, and we point to specific articles in the special issue that begin to answer these questions. We also look to literature from domains outside aviation psychology for guidance. Research on performance appraisal in the field of industrial psychology provides techniques and knowledge relevant to training instructors to evaluate crews reliably and validly. We conclude with a series of research questions that should be addressed.

Reference

Baker, D. P., & Dismukes, R. K. (in press). A framework for understanding crew performance assessment issues. *International Journal of Aviation Psychology.*

Pilot Instructor Rater Training:

The Utility of the Multifaceted IRT Model

Casey Mulqueen

David P. Baker

American Institutes for Research

R. Key Dismukes

National Aeronautical Space Administration

Ames Research Center

Abstract

A Multifaceted one-parameter item response theory (i.e., Rasch) model was used to examine inter-rater reliability training for pilot instructors. This model provides a means for examining individual instructor leniency or severity in ratings, difficulty of grade sheet items, skill levels of flight crews, and interactions among these components. It was found that pilot instructor trainees differed in their levels of rating severity and that higher CRM scores were easier to achieve than technical scores. Interaction analyses identified several pilot instructors who were evaluating crews in an unexpected manner, which is useful when providing feedback during training.

Reference

Mulqueen, C., Baker, D. P., & Dismukes, R. K. (in press). Pilot instructor training: The utility of the multifaceted IRT model. *International Journal of Aviation Psychology*.

Within Group-Versus Between-Group Consistency:

Examining the Effectiveness of IRR Training

David P. Baker

Casey Mulqueen

American Institutes for Research

R. Key Dismukes

NASA-Ames Research Center

Abstract

Inter-rater reliability (IRR) training has been proposed as an effective strategy for training pilot instructors to accurately assess crew performance. This training usually takes place during a one-day workshop in which pilot instructors watch and assess the videotaped performance of several crews flying scenarios or their component event sets. While reasonable levels of inter-rater agreement have been reported for IRR training, these results are typically reported at the within–group level. At large air carriers, where pilot instructor/evaluators are trained in numerous workshops, between-group agreement is equally important. This paper explores the extent to which between-group differences exist across several IRR classes.

Reference

Baker, D. P., Mulqueen, C., & Dismukes, R. K. (in press). Within versus between-group consistency: Examining the effectiveness of IRR training. *Proceedings of the Eleventh International Symposium on Aviation Psychology.*

Training Raters to Assess Resource Management Skills

David P. Baker

Casey Mulqueen

American Institutes for Research

R. Key Dismukes

NASA Ames Research Center

Introduction

Training work teams in resource management is becoming increasingly important in a wide variety of organizations and industries. For example, within commercial aviation, effective resource management is critical on the flight deck. In this industry, where the consequences of error are extreme, the vast majority of incidents and accidents have been attributed to breakdowns in the resource management skills of crew members (Helmreich, Foushee, Benson, & Russini, 1986; Helmreich, Weiner, & Kanki, 1993; Prince & Salas, 1993; Ruffell-Smith, 1979). As a result, commercial aviation has been a leading contributor in the development of effective resource management training, or crew resource management (CRM) as it is known in the airline industry. This training has continually evolved over the last twenty years from short lecture and discussion-based classes focused on aircrew members' attitudes toward teamwork to a fully integrated performance-based training curriculum known as the Advanced Qualification Program (AQP) (Birnback & Longridge, 1993).

An important feature of AQP is the fact that aircrew members must complete a Line Operation Evaluation (LOE) scenario at the end of initial and recurrent training. This type of training event is similar to other resource management training programs in which trainees are provided with practice and feedback or are evaluated at the end of training on their resource management skills (Baker & Salas, 1997; Brannick, Salas, & Prince, 1997). Essentially, an LOE is a job simulation that includes identifiable scenario events that are designed to elicit technical and CRM behaviors by the crew (ATA, 1994). A pilot instructor observes a crew's performance during the LOE and rates the crew on specific technical and CRM skills. These ratings are used to determine whether or not the pilots comprising the crew should be certified to fly the line or require additional training.

A critical factor in the evaluation a flight crew's resource management skills is the pilot instructor. Inevitably, the reliability and validity of the process rests upon the ability of the instructor to observe relevant crew behavior and make an accurate evaluation that is recorded on the rating form. As Birnbach and Longridge (1993) noted, LOEs can only be valid if pilot instructor ratings are accurate and reliable. The most direct and efficient method for ensuring that pilot instructors will be capable of evaluating a crew's resource management is to provide them with rater training. Formal rater training can serve to familiarize pilot instructors with the scenario events, the rating forms, and the CRM skills to be assessed.

The primary purpose of this chapter is to examine the relevant research literature on rater training in order to develop a series of guidelines for training raters to evaluate resource management skills. To do this, we will first examine how rater training is conducted in airlines, as well as review the available empirical literature on its effectiveness. As mentioned earlier, the commercial airline industry has been one of the leaders in all aspects of resource management training. Second, we will review four strategies that have been traditionally used to train supervisors who conduct performance appraisals. In addition, we will present research on each strategy's effectiveness and discuss the relative merits of these approaches. Finally, the results from the literature review will be combined and summarized into a set of guidelines for developing rater training in the future. These guidelines delineate what we believe are the "best practices" for training raters to assess resource management skills.

Reference

Baker, D. P., Mulqueen, C., & Dismukes, R. K. (2001). Training raters to assess resource management skills. In E. Salas, C. A. Bowers & E. Eden (Eds.), *Improving teamwork in organizations: Applications of resource management training* (pp. 131-145). Mahwah, NJ: Erlbaum.

Using Multifacet Rasch Analysis to Examine the Effectiveness of Rater Training

Casey Mulqueen

David Baker

American Institutes for Research

R. Key Dismukes

NASA Ames Research Center

Abstract

Multifacet Rasch (e.g., one-parameter IRT) analysis was used to examine the effectiveness of rater training for individuals that are required to conduct end-of-training work performance evaluations. The results are presented with emphasis on the additional information provided by this technique, and the relative advantages and disadvantages of this approach vis-a-vis other methods of analysis.

Reference

Mulqueen, C., Baker, D. P., & Dismukes, R. K. (2000, April). *Using multifacet Rasch analysis to examine the effectiveness of rater training.* Paper presented at the 15th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Assessing I/E Rater Training Effectiveness: Issues in Measurement

Casey Mulqueen

David P. Baker

American Institutes for Research

Washington, D.C.

Abstract

In order to achieve valid evaluations of flight crew performance, the ratings that are provided by I/Es must be reliable (Birnbach & Longridge, 1993). Reliability of the rating process can be strengthened through the use of rater training programs. An important follow-up of any rater training program is a formal assessment of its effectiveness. This paper will briefly explore various methods used to assess the effectiveness of I/E rater training, in particular methods for evaluating the reliability of ratings provided by the I/Es. A multifaceted methodology for assessing interrater reliability (IRR) will be described, along with an example of its use following an I/E rater training program.

Reference

Mulqueen, C. & Baker, D. P. (1999). Assessing I/E training effectiveness: Issues for measurement. *Proceedings of the Tenth International Symposium on Aviation Psychology*, *1*, 323-328.

Pilot Instructor/Evaluator Rater Training: Guidelines for Development

David P. Baker

Casey Mulqueen

American Institutes for Research

Washington, DC

Abstract

The extent to which pilot instructors are trained to reliably and accurately assess an aircrew's CRM and technical performance during a LOS scenario is critical under the Advanced Qualification Program. Pilot instructor/evaluators must be reliable and accurate to ensure that valid feedback is provided to aircrews undergoing training and that sound decisions are made regarding the certification of aircrews to fly the line. To address the critical need for effective rater training, this document reviews the relevant research on several strategies for training raters. Based on this review, a series of guidelines are presented for structuring pilot instructor/evaluator rater-training programs.

Reference

Baker, D. P., & Mulqueen, C. (1999). Pilot instructor/evaluator rater training: Guidelines for development. *Proceedings of the Tenth International Symposium on Aviation Psychology*, *1*, 332-337.

Training Pilot Instructors to Assess CRM:

The Utility of Frame-Of-Reference (FOR) Training

David P. Baker

Casey Mulqueen

American Institutes for Research

R. Key Dismukes

Aerospace Human Factors Research Division

NASA Ames Research Center

Abstract

The extent to which pilot instructors are trained to assess crew resource management (CRM) skills accurately during a simulator scenario is critical. Pilot instructors must make accurate performance ratings to ensure that proper feedback is provided to the flight crew and appropriate decisions are made regarding certification to fly the line. This paper reviews several approaches to rater training and identifies what we believe would be the most effective approach for training pilot instructors to assess CRM: Frame-Of-Reference (FOR) training. The goal of FOR training is to train pilot instructors to common standards, which are developed by expert instructors. Research suggests that if pilot instructors were trained to evaluate performance using the same standards as "experts" they should produce more accurate ratings. Based on the results of this research, specific guidelines are presented for developing FOR training and the benefits and limitations of this training are discussed. Finally, we conclude with a series of unanswered questions regarding pilot instructor rater training that require investigation within the airline industry.

Reference

Baker, D. P., Mulqueen, C., & Dismukes, R. K. (1999). Training pilot instructors to assess CRM: The utility of frame-of-reference (FOR) training. *Proceedings of the International Aviation Training Symposium*, 291-300.

Grant Publications and Presentations

Baker, D. P., & Dismukes, R. K. (submitted for publication). Gold standards training. *Handbook on Human Factors and Ergonomic Methods.*

Baker, D. P., & Dismukes, R. K. (in press). A Framework for Understanding Crew Performance Assessment Issues. *International Journal of Aviation Psychology.*

Baker, D. P., Mulqueen, C. & Dismukes, R. K. (in press). With-in versus between-group consistency: Examining the effectiveness of IRR training. *Proceedings of the Eleventh International Symposium on Aviation Psychology.*

Mulqueen, C., Baker, D. P., & Dismukes, R. K. (in press). Pilot instructor rater training: The utility of the multifaceted IRT model. *International Journal of Aviation Psychology.*

Baker, D. P., Mulqueen, C., & Dismukes, R. K. (2001). Training raters to assess resource management skills. To appear in E. Salas, C. Bowers & E. Edens (Eds.). *Applying Resource Management in Organizations: A guide for training professionals* (pp. 131-145). Mahwah, NJ: Lawrence Erlbaum Associates.

Mulqueen, C., & Baker, D. P., & Dismukes, R. K. (2000, April). *Using Multifacet rasch analysis to examine the effectiveness of rater training.* Paper presented at the 15th Annual Conference for the Society for Industrial and Organizational Psychology, New Orleans, LA.

Baker, D. P. (2000, January). *Future directions: Gold standards training.* Presentation at the 9th Annual Pilot Instructor Training Seminar, Aer Lingus, Dublin, Ireland.

Baker, D. P. (2000, January). *Behavioral observation training.* Presentation at the 9th Annual Pilot Instructor Training Seminar, Aer Lingus, Dublin, Ireland.

Baker, D. P., Mulqueen, C., & Dismukes, R. K. (1999). Training pilot instructors to assess CRM: The utility of frame-of-reference (FOR) training. *Proceedings of the International Aviation Training Symposium*, 291-300.

Baker, D. P., & Dismukes, R. K. (1999, May). *Training instructor/evaluators to assess CRM: From research to practice.* Symposium presented at the Tenth International Symposium on Aviation Psychology, Columbus, OH.

Baker, D. P., & Mulqueen, C. (1999). Pilot instructor/evaluator rater training: Guidelines for development. *Proceedings of the Tenth International Symposium on Aviation Psychology*, 332-337.

Mulqueen, C., & Baker, D. P. (1999). Assessing I/E training effectiveness: Issues for measurement. *Proceedings of the Tenth International Symposium on Aviation Psychology*, 323-328.

APPENDIX A: GOLD STANDARDS COURSE DESIGN GUIDE

AGENDA

8:00 - 8:30	INTRODUCTION
8:30 - 9:30	USING LOE GRADE SHEETS
9:30 - 10:30	REPEATING EVENT SETS
10:30 – 12:30	PRACTICE VIDEOTAPES
12:30 - 13:00	LUNCH BREAK
13:30 - 14:00	BEHAVIORAL OBSERVATION TRAINING
14:00 - 16:30	GOLD STANDARDS TRAINING AND POST-TRAINING EXERCISE

COURSE DESCRIPTION AND OVERVIEW

Course: Introduction

Instructional Objectives: 1.A through 1.C

Time: 8:00 - 8:30

Description

This module provides new pilot instructor/evaluators (I/Es) with general background information regarding the role of I/Es, the role of performance ratings in the Advanced Qualification Program (AQP), and the objectives of Gold Standard training. Emphasis is placed on the importance of quality ratings so that carrier management can make well-informed decisions regarding crew training and operational safety.

Upon completing this module, trainees will be able to:

- describe the role of I/Es;
- describe the role of performance ratings in AQP; and
- describe the objectives of Gold Standards training.

MAJOR POINTS	ENABLING OBJECTIVES
✈ Describe the various tasks that trainees are likely to perform as I/Es. Describe the responsibilities they are expected to provide and the types of evaluations that they are responsible for administering line operational evaluations (LOEs). Emphasize the importance of performance feedback as a mechanism for changing pilots' attitudes and behavior.	A.1
✈ Describe AQP, the concept of proficiency-based training, and the use of LOE in AQP. State that data collected during the LOE are analyzed for trends across fleets, within fleets, and across time. Emphasize that the results of these analyses are used to revise AQP training curricula in an iterative fashion.	B.1
✈ Provide specific examples of how topic grades, event set grades, and overall grades for the LOE can be used to make operational decisions regarding safety and training. Example: LOE grades can be used to assess pilot proficiency on different maneuvers. If performance drops below some minimum level, special purpose training can be developed to address the problem.	B.2
✈ Describe how the Gold Standards represent the judgment of expert I/Es. Describe how Gold Standards will help new I/Es adopt a common frame of reference when evaluating crews in the simulator.	C.1
✈ Describe the mechanics of Gold Standards training. Emphasize that new I/Es will practice and receive feedback regarding how to complete LOE grade sheets, how to perform repeats, and how to evaluate crew performance. Emphasize that their training will involve verbal instruction, practice exercises, and group discussion.	C.2

COURSE: INTRODUCTION

OBJECTIVE 1.A: To enable trainees to describe the role of pilot instructor/evaluators.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
1.A.1) Describe the major tasks required of I/Es (e.g., PIs, SCs) in the LOE process.	Tutorial	Overheads	Oral	Main tasks include: 1. Manipulating the simulator controls. 2. Interacting with crews by role-playing the ATC. 3. Evaluating crew performance. 4. Providing performance-based feedback.	Knowledge

***Presented in order of importance.**

COURSE: INTRODUCTION

OBJECTIVE 1.B: To enable trainees to describe the uses of performance ratings in AQP.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
1.B.1) Describe how performance ratings fit in the AQP model of training and evaluation.	Tutorial	Overheads	Oral	AQP is a proficiency-based training program. LOE grades are analyzed for trends. This information is used to revise curricula in an iterative fashion. Result: crew performance ratings allow carrier management to make informed decisions about training issues.	Knowledge
1.B.2) Describe specific uses of topic grades, event set grades, and overall grades.	Tutorial	Overheads	Oral	Topic grades, event set grades, and LOE overall grades are used to: 1. Provide assurances of proficiency levels. 2. Validate training assumptions. 3. Analyze the effectiveness of AQP training. 4. Provide performance feedback. 5. Refine the training and measurement processes.	Knowledge

*Presented in order of importance.

COURSE: INTRODUCTION

OBJECTIVE 1.C: To enable trainees to describe the goals of Gold Standards training.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
1.C.1) Describe how Gold Standards training will calibrate all new I/Es to a common frame of reference.	Tutorial	Overhead	Oral	Gold Standards are based on the judgments of expert instructor/ evaluators. They represent the carrier's definition of what constitutes acceptable/unacceptable crew performance. The objective of Gold Standards training is to calibrate new I/Es to this common frame of reference.	Knowledge
1.C.2) Provide an overview of Gold Standards training.	Tutorial	Overhead	Oral	First, trainees will learn how to complete LOE worksheets. Next, they will learn how to repeat event sets (when necessary). Finally, they will make practice ratings of crew performance using videotaped examples. Feedback will be provided regarding discrepancies between their individual ratings and the Gold Standards. The rationale for these discrepancies will be discussed in detail.	Knowledge

*** Presented in order of importance.**

Course Description and Overview

Course: Using LOE Grade Sheets

Instructional Objectives: 2.A through 2.E

Time: 8:30 - 9:30

Description

This module provides instruction on how to complete LOE grade sheets. Emphasis is placed on understanding scale definitions and aggregating topic grades to create overall ratings of crew performance.

Upon completing this module, trainees will be able to:

- describe the scales used for CRM and TECH topics, CRM and TECH event set grades, and pilot-in-command (PIC) and second-in-command (SIC) overall grades;
- describe the process by which topic grades are translated into TECH and CRM event set grades;
- describe the process by which TECH and CRM topic and event set grades are translated into PIC and SIC overall grades; and
- describe the general criteria for success and failure in LOE.

MAJOR POINTS	ENABLING OBJECTIVES
✈ Describe the differences between CRM and TECH topic grades, CRM and TECH event set grades, and PIC and SIC overall grades. CRM and TECH topic grades refer to broad classes of behavior that can be directly observed. CRM and TECH event set grades refer to ratings of crew performance that are based upon the crewmembers' performance across topics for an event set. These grades are created using the success criteria that are listed on each grade sheet. PIC and SIC overall grades are ratings of each individual crewmembers' performance throughout the event set. These grades are based upon the CRM and TECH topic and event set grades plus the I/E's judgment.	A.1
✈ Describe the scale that is used to grade CRM topics. Emphasize that a "Missed observation" means that the I/E did not see the behavior for a reason unrelated to the crew's performance, such as being distracted while manipulating the simulator controls. This is not to be confused with "Not performed" which refers to specific CRM topics that the crewmembers failed to perform.	A.2
✈ Describe the scale that is used to grade TECH topics. Emphasize that a grade of "1" (Repeat) for a TECH topic does not require a repeat.	A.3
✈ Describe the scales that are used to grade CRM and TECH event set performance. Again, emphasize that a grades of "1" (Repeat) do not require a repeat.	A.4
✈ Describe the scales that are used to grade PIC and SIC overall performance on the event set. Point out that a value of "1" (Repeat) for the PIC or SIC requires a repeat of the event set or parts thereof.	A.5

MAJOR POINTS	ENABLING OBJECTIVES
✈ Describe how to grade crew performance on CRM and TECH topics. Emphasize that the crews should demonstrate knowledge of relevant SOP and flight manuals. Also note that the aircraft must be operated within standards.	B.1
✈ Point out the success criteria at the bottom of each LOE grade sheet. Emphasize that these criteria provide explicit instructions for determining CRM and TECH event set grades, and that they may vary across event sets. Example: CRM performance for the event set is graded as "1" if three or more CRM topics are checked as "Not Performed". Example: TECH performance for the event set is graded as "1" if two or more TECH topics are graded as less than "Standard" <u>or</u> any TECH topic is graded as repeat.	B.2
✈ Emphasize that PIC and SIC overall grades are to be based on the crewmembers' behavior during the event set. This is typically done by considering the crew's overall CRM and TECH proficiency coupled with the I/E's judgment.	C.1
✈ Describe the relative importance of CRM and TECH behaviors when determining PIC and SIC overall grades. Note that PIC and SIC grades must be based on proficiency objectives and not solely on CRM performance.	C.2
✈ Describe how supporting comments are always important. However, stress that supporting comments are <u>absolutely required</u> for grades of "repeat" (1), "debriefed" (2), and "excellent" (4). Note that these grades are used by management to better understand performance trends in the AQP.	C.3
✈ Describe the general criteria for LOE success. Emphasize that these guidelines are meant to supplement, not replace, the topic ratings.	D.1
✈ Describe the criteria that lead to automatic ratings of "Unsatisfactory" overall performance on the LOE. Note how these criteria work in conjunction with the general success criteria to assist I/Es in their task.	D.2

COURSE: USING LOE GRADE SHEETS

OBJECTIVE 2.A: To enable trainees to describe the scales used to assign CRM and TECH topic grades, CRM and TECH event set grades, and PIC and SIC overall grades.

***Enabling Objectives**	**Strategy**	**Media**	**Evaluation**	**Instructional Content**	**Type of Learning**
2.A.1) Identify the three major types of grades for each event set.	Tutorial	Overheads	Oral	Grades are assigned for: 1. CRM and TECH topics 2. Overall CRM and TECH for each event set 3. Overall PIC and SIC performance on the event set	Knowledge
2.A.2) Describe the scale that is used for grading CRM topics.	Tutorial	Overheads	Oral	CRM topics are graded as: 1. Missed observation 2. Not performed 3. Partially performed 4. Performed	Knowledge
2. A.3) Describe the scale that is used for grading TECH topics.	Tutorial	Overheads	Oral	TECH topics are graded as: 1. Repeat 2. Debriefed 3. Standard. 4. Excellent Note that a rating of "1" (Repeat) does <u>not require</u> the crew to repeat the event set.	Knowledge

***Presented in order of importance.**

COURSE: USING LOE GRADE SHEETS

OBJECTIVE 2.A: To enable trainees to describe the scales used to assign CRM and TECH topic grades, CRM and TECH event set grades, and PIC and SIC overall grades.

***Enabling Objectives**	**Strategy**	**Media**	**Evaluation**	**Instructional Content**	**Type of Learning**
2.A.4) Describe the scales that are used for grading CRM and TECH event set performance.	Tutorial	Overheads	Oral	CRM and TECH event set performance are graded as: 1. Repeat 2. Debriefed 3. Standard 4. Excellent Note that a rating of "1" (Repeat) does not require the crew to repeat the event set.	Knowledge
2.A.5) Describe the scales that are used for grading overall PIC and SIC performance on an event set.	Tutorial	Overheads	Oral	PIC and SIC performance are graded as: 1. Repeat 2. Debriefed 3. Standard 4. Excellent Note that a rating of "1" (Repeat) requires the crew to repeat the event set.	Knowledge

***Presented in order of importance.**

COURSE: USING LOE GRADE SHEETS

OBJECTIVE 2.B: To enable trainees to describe the process by which topic grades are translated into TECH and CRM event set grades.

***Enabling Objectives**	**Strategy**	**Media**	**Evaluation**	**Instructional Content**	**Type of Learning**
2.B.1) Describe how to evaluate performance on CRM and TECH topics.	Tutorial	Overheads	Oral	Crew should demonstrate knowledge of carrier SOP and comply with procedures in the FM and FOM. The aircraft should be operated within qualification standards.	Knowledge
2.B.2) Describe how CRM and TECH event set grades are computed.	Tutorial	Overheads	Oral	CRM and TECH event set grades are calculated using success criteria listed on the grade sheet. There are separate success criteria for CRM and TECH event set grades. There are also separate success criteria for each event set.	Knowledge

*** Presented in order of importance.**

COURSE: USING LOE GRADE SHEETS

OBJECTIVE 2.C: To enable trainees to describe the process by which topic and event set grades are translated into PIC and SIC grades.

***Enabling Objectives**	**Strategy**	**Media**	**Evaluation**	**Instructional Content**	**Type of Learning**
2.C.1) Describe how overall PIC and SIC grades are computed.	Tutorial	Overheads	Oral	PIC and SIC grades are calculated using topic and event set grades coupled with the I/E's judgment.	Knowledge
2.C.2) Describe the importance of CRM in determining overall PIC and SIC grades.	Tutorial	Overheads	Oral	The overall PIC and SIC grades must be based on general or specific proficiency objectives. They may not be based solely on CRM performance.	Knowledge
2.C.3) Describe the role of supporting comments.	Tutorial	Overheads	Oral	Comments are included in an AQP database along with crewmembers' grades. These comments help management understand the meaning behind the grades assigned. Further, comments suggest areas for improving the training program. Supporting comments are always important, and should be included as often as possible. However, ratings of "repeat" (1), "debriefed" (2) and "excellent" (4) absolutely require supporting comments.	Knowledge

*** Presented in order of importance.**

COURSE: USING LOE GRADE SHEETS

OBJECTIVE 2.D: To enable trainees to describe the general LOE criteria for success and failure.

***Enabling Objectives**	**Strategy**	**Media**	**Evaluation**	**Instructional Content**	**Type of Learning**
2.D.1) Identify and describe the general criteria for success.	Tutorial	Overheads	Oral	The general criteria for success in the LOE are: 1. The aircraft landed safely. 2. The flight flew within legal limits with momentary deviations. 3. The flight remained within SOP or deviations were justified. 4. Appropriate action was taken in a timely manner. 5. All event sets were graded "Excellent", "Standard" or "Debriefed" by conclusion of LOE.	Knowledge
2.D.2) Identify and describe factors that are considered unsatisfactory.	Tutorial	Overheads	Oral	An LOE is considered unsatisfactory if: 1. A repeated event set is not rated as "Debrief" or higher. 2. The crew receives a "Repeat" on three event sets. 3. The crew crashes the simulator. 4. The crew performs a gross deviation in a single event set that compromises the aircraft to the point of an imminent crash.	Knowledge

***Presented in order of importance.**

COURSE DESCRIPTION AND OVERVIEW

Course: Repeating Event Sets

Instructional Objectives: 3.A through 3.B

Time: 9:30 - 10:30

Description

This module provides instruction on repeating event sets. Emphasis is placed on specific strategies for repeating event sets, "alternative" repeat procedures, and tips for identifying problem scenarios. Group discussions and short exercises provide trainees with hands-on practice.

Upon completing this module, trainees will be able to:

- ✈ identify, describe, and apply specific strategies for repeating event sets; and
- ✈ select and execute an appropriate strategy for repeating event sets given time, resource, and other constraints.

MAJOR POINTS	ENABLING OBJECTIVES
✈ Describe when an event set needs to be repeated. Emphasize that only PIC and SIC overall grades of "Repeat" (1) actually need to be repeated. TECH topics and CRM or TECH event set grades of "Repeat" (1) do not need to be repeated.	A.1
✈ Describe the rationale behind allowing event sets to "play themselves out" to a logical conclusion. Emphasize that this technique allows crewmembers to observe the effects of their behavior (e.g., "error chains") in a safe, natural environment. Ask the trainees for examples based on their own experiences.	A.2
✈ Describe the rationale behind not debriefing the crew until the LOE is complete. Emphasize that to avoid compromising the learning experience, crewmembers should not be coached regarding their performance, SOP, or situational cues. The I/E should only specify which event set requires repeating.	A.3
✈ Describe specific strategies for repeating an event set using examples based on case studies.	A.4
✈ Provide the trainees with case studies that may require repeating an event set. Ask trainees to provide possible repeat strategies for each case study. Example: Using script-based clues to modify the event set based on the crews' prior behavior. Example: Orally quizzing the crew on specific facts regarding the relevant system, maneuver, or procedure.	A.5

COURSE: REPEATING EVENT SETS

OBJECTIVE 3.A: To enable trainees to identify, describe, and apply specific strategies for repeating event sets.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
3.A.1) Identify when a repeat is required.	Tutorial	Overheads	Oral	Repeats are only required for PIC and SIC overall grades of "Repeat" (1). Technical topics, CRM and TECH event set grades of "Repeat" (1) do <u>not</u> need to be repeated.	Knowledge
3.A.2) Describe the importance of and rationale behind allowing event sets to play themselves out to a logical conclusion.	Tutorial	Overheads	Oral	Allowing event sets to play themselves out permits crewmembers to observe the effect of their action/inaction on system performance in a safe learning environment.	Knowledge
3.A.3) Describe the importance of not debriefing the crew until the LOE is complete.	Tutorial	Overheads	Oral	The learning value of the repeat is compromised if crewmembers receive coaching on the success criteria. The instructor should only specify which phase of flight requires a repeat.	Knowledge

***Presented in order of importance.**

COURSE: REPEATING EVENT SETS

OBJECTIVE 3.A: To enable trainees to identify, describe, and apply specific strategies for repeating event sets.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
3.A.4) Present specific strategies for repeating an event set.	Tutorial	Overheads	Oral	Event sets can be repeated by: 1. Using script-based cues to modify the event set based on the crewmembers' prior decisions and/or behaviors. 2. Re-positioning the simulator. 3. Selecting a different event set of equal difficulty. 4. Quizzing the crew regarding the relevant system or procedure.	Knowledge
3.A.5) Apply specific strategies for repeating an event set.	Case studies (Appendix A)	Overheads/ Handouts	Oral	Use case studies to elicit group input and discussion for applying these strategies in the simulator. When considering trainees' suggestions, remember that repeats should be perceived as realistic, and as part of an uninterrupted flight. Also, various constraints, such as time and phase-of-flight may limit the alternatives that are available to the pilot instructor.	Skill

***Presented in order of importance.**

COURSE DESCRIPTION AND OVERVIEW

Course:	LOE Grading Practice
Instructional Objectives:	4.A through 4.B
Time:	10:30 - 12:30

Description

This module provides new I/Es opportunities to practice grading crew performance on LOEs. Emphasis is placed on understanding the behavioral dimensions and grading scale anchors prior to observing examples of crew performance. Practice ratings are made using videotaped scenarios of crews performing in a full-motion simulator.

Upon completing this module, trainees will be able to:

- ✈ describe the skills that are being assessed in the LOE; and
- ✈ grade crews using the LOE grade sheet.

MAJOR POINTS	ENABLING OBJECTIVES
✈ For each videotaped event set, set the stage by describing the tasks that the crews are expected to perform. Next, describe the grade sheet that will be used to evaluate the crewmembers' performance. Provide specific examples of performance at the various levels on the grade sheet. Provide the rationale behind each performance level, using relevant SOP and FARs to support your position.	A.1
✈ Allow the trainees to practice rating the videotaped event sets. The practice videotape should include crews at varying levels of proficiency performing multiple event sets.	B.1

COURSE: LOE GRADING PRACTICE

OBJECTIVE 4.A: To enable trainees to describe the skills that are being evaluated in the LOE.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
4.A.1) Describe the tasks to be performed during the event set and the scales that will be used to assess the crewmembers' performance.	Tutorial	Overheads	Oral	Define the tasks that are to be performed in each event set. Use concrete examples as necessary. Describe the scales that will be used to assess the crewmembers' performance. Provide examples of performance at various levels of proficiency for each scale on the LOE grade sheet. This should require between 30 and 45 minutes to complete.	Knowledge

***Presented in order of importance.**

COURSE: LOE GRADING PRACTICE

OBJECTIVE 4.B: To enable trainees to grade crews using the LOE grade sheet.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
4.B.1) Practice rating videos of crews flying LOE event sets.	Practice	Videotaped scenarios	Oral	After describing the dimensions and the scale anchors, allow the trainees to practice rating videotaped scenarios of crew performance. The videotape (approximately 60-80 minutes in length) should include crews at varying levels of proficiency performing multiple event sets.	Skill

***Presented in order of importance.**

Course Description and Overview

Course: Behavioral Observation Training

Instructional Objectives: 5.A through 5.B

Time: 13:30 - 14:00

Description

This module provides instruction on improving new I/Es' observation skills. Emphasis is placed on distinguishing between descriptions of behavior and conclusions regarding the effectiveness of those behaviors. Several strategies are presented for improving the trainees' observational skills. This module is conducted while trained support staff are analyzing the performance ratings from the previous module (LOE Grading Practice).

Upon completing this module, trainees will be able to:

- distinguish between behaviors and conclusions; and
- identify and describe five guidelines for effective observation.

MAJOR POINTS	**ENABLING OBJECTIVES**
✈ Describe the distinction between descriptions of crew behavior and conclusions regarding the effectiveness of those behaviors. Remind the trainees that behavioral descriptions refer to specific, discrete tasks that were or were not performed by the crew. Conclusions, on the other hand, refer to inferences and judgments made by the pilot instructor. As a result, they are more subject to perceptual biases.	A.1
✈ Describe five guidelines for effective behavioral observation. Note how these guidelines should be used when making notes in the "comments" section of the LOE worksheet (Objective 2.C).	B.1

COURSE: BEHAVIORAL OBSERVATION TRAINING

OBJECTIVE 5.A: To enable trainees to distinguish between behaviors and conclusions.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
5.A.1) Describe the distinction between descriptions of behavior and conclusions regarding the effectiveness of those behaviors.	Tutorial (Appendix B)	Overheads/ Handouts	Written	Behavioral descriptions provide crews with feedback regarding actions that were or were not taken by the crew. Conclusions regarding behavior are usually based on and I/E's assumptions of what the crewmembers may or may not have been thinking. As a result, they are subject to bias and misinterpretation.	Knowledge

***Presented in order of importance.**

COURSE: BEHAVIORAL OBSERVATION TRAINING

OBJECTIVE 5.B: To enable trainees to identify and describe five guidelines for effective observation.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
5.B.1) Identify and describe five guidelines for effective behavioral observation.	Tutorial/ Group exercise (Appendix C)	Overheads	Oral	Guidelines for effective observation include: 1. Use specific examples. 2. Avoid adjective qualifiers. 3. Avoid assumptions about crewmembers' knowledge. 4. Avoid the use of quantitative values. 5. Provide enough detail to determine the extent of situational effects. Emphasize that these guidelines can be helpful when making comments regarding the crew's performance (Objective 2.C).	Knowledge

*** Presented in order of importance.**

COURSE DESCRIPTION AND OVERVIEW

Course: Gold Standards Training and Post-Training Videotape

Instructional Objectives: 6.A - 6.B

Time: 14:00 - 16:30

Description

This module provides feedback that compares the each new I/E's practice ratings with the Gold Standards. Group discussion is used to explain the rationale behind the Gold Standard ratings, and to solidify the decision rules that were specified in Module 4 “Evaluating Crew Performance with Gold Standards.” I/Es grade a post-training videotape to evaluate the extent to which trainees have improved their skills.

Upon completing this module, trainees will be able to:

- ✈ interpret the degree of similarity between their individual ratings and the Gold Standards; and
- ✈ interpret their level of skill acquisition as a result of training.

MAJOR POINTS	ENABLING OBJECTIVES
✈ Describe how Gold Standards will be used to calibrate all new I/Es using a common frame-of-reference. Emphasize that Gold Standards training was developed to ensure that all crewmembers will be evaluated consistently, regardless of which I/E evaluates them.	A.1
✈ Describe how the Gold Standards represent the ratings of a panel of expert I/Es. Emphasize that the Gold Standards are based on carrier SOP and relevant FARs.	A.2
✈ Describe the concept of "deviation scores" as the difference between a given I/E's rating and the gold standard. Emphasize that because these are "deviations," lower scores are better, with perfect agreement to the Gold Standard being equal to zero.	A.3
✈ Provide feedback on an item-by-item basis. Identify the rationale for discrepancies between individual ratings and the Gold Standards. Consult relevant FARs and carrier SOP to identify why the discrepancies occurred, so that I/Es leave training with a common frame-of-reference.	A.4
✈ Have I/Es grade post-training videotape. Compare pre- and post-training performance as an indicator of skill improvement. Provide feedback to individual trainees at a later time (e.g., via e-mail) to help them gauge their level of skill acquisition.	B.1

COURSE: GOLD STANDARDS TRAINING AND POST-TRAINING VIDEOTAPE

OBJECTIVE 6.A: To enable trainees to interpret the degree of similarity between their individual ratings and the Gold Standards.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
6.A.1) Describe the purpose of rater calibration using Gold Standards.	Tutorial	Overheads	Oral	The purpose is to ensure that there are no systematic differences among raters.	Knowledge
6.A.2) Describe the process by which gold standard ratings were developed.	Tutorial	Overheads	Oral	Groups of expert I/Es convened to rate the videotaped examples and discuss the rationale for their ratings. Relevant FARs and SOP were consulted for support. Final ratings represent consensus among these experts.	Knowledge
6.A.3) Describe the concept of "deviation scores."	Tutorial	Overheads	Oral	Deviation scores reflect the difference between the pilot instructor's rating and that of the Gold Standard. Higher values indicate greater disagreement. Perfect agreement = 0.0	Knowledge
6.A.4) Provide feedback to trainees regarding their performance.	Tutorial	Overheads	Written	Feedback is presented on an item-by-item basis, with an emphasis on items that showed low agreement.	Knowledge
6.A.5) Identify the rationale for the observed discrepancy (if any).	Group Discussion	Overheads	Oral	Gold Standard ratings are based on FARs and carrier SOP. Identify discrepancies between individual ratings and the gold standard, and provide supporting evidence. Solicit group discussion to clarify issues.	Knowledge

***Presented in order of importance.**

COURSE: GOLD STANDARDS TRAINING AND POST-TRAINING VIDEOTAPE

OBJECTIVE 6.B: To enable trainees to interpret their level of skill acquisition as a result of training.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
6.B.1) Ascertain trainees' level of skill acquisition via a post-training exercise.	Practice	Videotaped scenarios	Written	The purpose is to determine the extent to which skills have improved as a result of training. I/E trainees will rate a new videotape in the same manner as before. Comparison of pre- and post-training performance will be used as an indicator of skill improvement. Due to time constraints, feedback will not be provided to the group as a whole. Rather, feedback will be provided to trainees at a later time (e.g., via e-mail) to help them gauge their individual progress.	Skill

***Presented in order of importance.**