NASA/TM—2003–212809

A Gold Standards Approach to Training Instructors to Evaluate Crew Performance

David P. Baker and R. Key Dismukes
Ames Research Center, Moffett Field, California

December 2003

The NASA STI Program Office ... in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results ... even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at ***http://www.sti.nasa.gov***

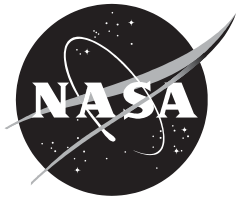
- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA STI Help Desk at (301) 621-0134

- Telephone the NASA STI Help Desk at (301) 621-0390

- Write to:
  NASA STI Help Desk
  NASA Center for AeroSpace Information
  7121 Standard Drive
  Hanover, MD 21076-1320

NASA/TM—2003–212809



A Gold Standards Approach to Training Instructors to Evaluate Crew Performance

David P. Baker
American Institutes for Research, Washington, DC

R. Key Dismukes
Ames Research Center, Moffett Field, California

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

December 2003

Acknowledgments

Funding for this work was provided by NASA's Aviation Safety Program.

Available from:

NASA Center for AeroSpace Information
7121 Standard Drive
Hanover, MD 21076-1320
301-621-0390

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
703-605-6000

This report is also available in electronic form at http://human-factors.arc.nasa.gov/ihs/flightcognition/

A Gold Standards Approach to Training Instructors to Evaluate Crew Performance[1]

David P. Baker
American Institutes for Research

R. Key Dismukes
NASA Ames Research Center

Funding for this work was provided by NASA's Aviation Safety Program

Introduction

The Advanced Qualification Program requires that airlines evaluate crew performance in Line Oriented Simulation. For this evaluation to be meaningful, instructors must observe relevant crew behaviors and evaluate those behaviors consistently and accurately against standards established by the airline. The airline industry has largely settled on an approach in which instructors evaluate crew performance on a series of event sets, using standardized grade sheets on which behaviors specific to event set are listed. Typically, new instructors are given a class in which they learn to use the grade sheets and practice evaluating crew performance observed on videotapes. These classes emphasize reliability, providing detailed instruction and practice in scoring so that all instructors within a given class will give similar scores to similar performance.

Only a few studies have examined the reliability achieved in typical classes for new instructors, however, the limited data available suggest that it can be fairly good: instructors within a given class give fairly consistent ratings (Baker, Mulqueen, & Dismukes, in press; Goldsmith & Johnson, in press; Holt, Hansberger, & Boehm-Davis, in press). However, the existing approach has important limitations; (1) ratings within one class of new instructors may differ from those of other classes; (2) ratings may not be driven primarily by the specific behaviors on which the company wanted the crews to be scored; and (3) ratings may not be calibrated to company standards for level of performance skill required. In this paper we provide a method we have developed to extend the existing method of training instructors to address these three limitations. We call this method the "gold standards" approach because it uses ratings from the company's most experienced instructors as the basis for training rater accuracy. Further, this approach ties the training to the specific behaviors on which the experienced instructors based their ratings. Gold standards training focuses on teaching new instructors to rate crew performance the same way highly experienced instructors do.

[1] This research was supported by a grant from the NASA Ames Research Center. The views presented in this paper are those of the author(s) and should not be construed as an official NASA position, policy or decision, unless so designated by other official document.

The gold standards approach is based on preparing annotated videotapes of crews performing at several levels of effectiveness in specific event sets.[2] The airline-training department assembles a team of highly experienced instructors who view the videotapes and identify strong points and weak points of crew performance relevant to the skills on which the crew is to be evaluated. Through discussion the instructors reach consensus on what grade to give for each event set and which behaviors are relevant to that grade. These grades and behaviors are listed in the annotation of the videotapes. During class new instructors can compare their ratings to the consensus ratings of experienced instructors and can discover on which specific behaviors the ratings should be based. Research has shown that formal evaluation of performance is most effective when evaluators are trained to conduct evaluation as a two-part process: (1) identification and observation of relevant behaviors and (2) scoring the relevant behaviors. The gold standards approach delineates these two aspects and provides training in both. In this paper we provide a practical description of how to use the gold standards approach.[3]

Gold Standards Training

In this section we provide an overview of each of the five modules that comprise Gold standards training. For each module, we describe the overall objectives, summarize the content, issues, and processes to be covered during the instruction, and describe strategies for reinforcing the learning objectives. Gold standards training incorporate much of material already used in airline classes for new instructors; thus it is a modification and extension of the existing approach rather than a replacement. The appendix provides an example of a syllabus for a class using the gold standards approach to train new instructors to evaluate crew performance. This class can be conducted in a single day in a typical airline-training department. It should be noted that this syllabus was developed for a specific air carrier therefore it may need to be modified if implemented at a different airline.

Module 1. Introduction

The first module should begin with an introduction that provides the new pilot instructor (PI)[4] with general background information regarding the role of a PI, the role of performance ratings in the Advanced Qualification Program (AQP), and the objectives of Gold Standard Training. The class leader should emphasize the importance of

[2] Baker, Swezey, & Dismukes (1980) describe specific methods for developing gold standards from videotapes.
[3] For discussion of the research foundation for the gold standards approach, see (Baker, Mulqueen & Dismukes, 2001).
[4] For convenience we refer only to pilot instructors, but this training is appropriate for any qualified individual directly involved in training and evaluating aircrew performance in LOS: instructors, check airmen, standards captains, etc.

accurate ratings to enable company management to make well-informed decisions regarding aircrew training and operational safety.

The introduction should inform PIs that they will be responsible for manipulating the simulator controls, interacting with crews by role-playing, evaluating aircrew performance, and providing performance-based feedback. The different types of assessments that the PIs will be responsible for making when evaluating crew performance should be introduced. Typically, PIs evaluate crews on three different levels of performance in accordance with predetermined criteria: observable behaviors associated with CRM and technical skills, performance on each event set that comprise the scenario (e.g., take-off, descent, landing), and overall performance on the scenario.

The concept of "gold standards" training is also introduced. Gold standards are based on the judgments of expert PIs and represent the carrier's definition of what constitutes acceptable/unacceptable aircrew performance. The objective of Gold standards training is to calibrate new PIs to this common frame of reference to ensure all crewmembers will be evaluated consistently, regardless of which PI evaluates them. It is important that the PI trainees understand the importance of reliably making ratings that are consistent with those of the expert PIs because the ratings are used to provide assurances of crew proficiency levels, provide crews with performance feedback, validate training assumptions, and refine the training content and evaluation tools.

This module ends with a review of the agenda for the remaining Gold standards training program. The trainees are informed that they will be instructed on required knowledge and procedures, given opportunities to practice applying the acquired information and skills, and provided feedback regarding their consistency and reliability in evaluating aircrew performance.

Module 2. Review performance standards

During the second module, PIs are instructed on the different types of assessments they will be required to make, the grade sheets they will be using, and relevant performance standards that will be used in evaluation. This module is the first step in developing consistent standards across PIs. Emphasis is placed on understanding grade sheet definitions and how to aggregate behavioral observations of CRM and technical skills into event set and overall scenario grades.

Reviewing the carrier's grade sheet should begin with a description of the different types of ratings that are required. PIs should be instructed on how to distinguish observable behaviors for both CRM and technical skills and how to rate each event set and crew performance on the scenario. Part of this general review should include an explanation of the different scales used to make different ratings. The review should explain the meaning attached to each of the points on each of the grade sheet rating scales. Any deviations in the meaning attached to the scale points across the scales used for the different types of ratings should also be explained and clarified with an example. Finally, any distinctions among the scale points within a rating scale should be

discussed. For example, a rating of “1” on the scale for evaluating observed CRM behaviors may represent “Missed Observation” indicating that the PI did not see the behavior for reasons unrelated to the air crew’s performance (e.g., PI being distracted when manipulating the simulator controls); whereas, a rating of “2” on the same scale may represent “Not Performed” indicating that the aircrew failed to perform the behavior.

Next, this module should include a detailed review of the carrier’s grade sheet. This requires a review of the scenario(s), the event sets, and the CRM and technical skills to be evaluated. For example, information regarding the requirements for successful aircrew performance on each event set within a particular scenario should be covered. If there are any pre-established grading rules that are used by the carrier (e.g., cases where certain behavioral observations lead to specific air crew performance ratings), these rules should be reviewed and discussed.

Module 3. Observation skills training

To ensure PIs consistently observe the relevant behaviors of aircrews during LOS, new PIs must receive instruction on how to accurately observe an aircrew’s performance during a scenario — Behavioral Observation Training (BOT). The third module is devoted to assisting PI trainees in differentiating between observation and evaluation processes. Observation processes involve the detection, perception, and recall of behavioral events; whereas, evaluation processes require the categorization, integration, and evaluation of what was observed.

BOT should include a discussion of the difference between descriptions of aircrew behavior versus conclusions regarding the effectiveness of those behaviors. PIs should be told that behavioral descriptions are specific, verifiable, discrete tasks that were or were not performed by the aircrew. These behavioral descriptions will be used to provide feedback regarding actions that a crew did or did not take. Conclusions are evaluations or judgments made by PIs regarding the effectiveness of the behaviors that the aircrew performed or failed to perform.

PIs should be then instructed on how to effectively document behavioral observations. Guidelines should be presented for effectively documenting behavioral descriptions: using specific examples, avoiding adjective qualifiers, avoiding assumptions about crewmembers’ knowledge, avoiding the use of quantitative values, and providing enough detail to determine the extent of situational effects.

To reinforce the instruction on accurately observing and documenting aircrew behaviors, PIs should be given practice opportunities with feedback. For example, PIs might watch a videotape of an aircrew performing a scenario and document the behaviors they observe the aircrew perform or fail to perform. The videotapes should be annotated with details regarding expert observations about the specific behaviors exhibited by the aircrew and how those behaviors are best interpreted. This annotation provides detailed feedback to the PIs so they can compare what they observed or failed

to observe and how they interpreted their observations to observations and interpretations of experts.

Module 4. Rating practice

This module provides the PI trainees with an opportunity to practice evaluating and grading aircrew performance during LOS. Specifically, PIs watch videotapes of aircrews performing event sets within different scenario(s) and then make ratings using company grade sheets that identify the behavior to be evaluated on each scenario. Ideally, the practice videotapes should portray scenarios to be used in actual LOFTs or LOEs the airline will be using to evaluate crews in the coming months so that the new PIs are exposed to the specific scenarios in which they will be evaluating crews. Practice videotapes should be selected to display a full range of aircrew performance — a minimum of three practice videotapes displaying excellent, average, and poor performance on each event set is recommended.

For each videotaped event set, it is important to first set the stage by describing the tasks the aircrews are expected to perform, using concrete examples when necessary, and reviewing the grade sheet and scales to be used to evaluate the aircrews' performance. Also, as part of the scale review, specific examples of performance at various levels of proficiency for each scale on the grade sheet should be discussed. The rationales for categorizing the examples at specific performance levels should be discussed, using relevant carrier SOP and FARs to support the categorizations.

Once the PIs have been briefed on the expected aircrew tasks, grading sheets, and rating scales for each videotaped event set, PIs should view the practice videotapes, document behavioral observations, and evaluate aircrew performance using the grading sheets provided. The practice videotapes are shown in a continuous manner. At the end of the practice session, grading sheets are collected.

Analyze rating data

Instructors must analyze the practice ratings from Module 4 and prepare materials for providing feedback during Module 5.

Table 1. Gold standard example.

SCENARIO EVENT SET 3

TRIGGER: System malfunction during climb-out: 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.

At that heart of these analyses is the comparison of the PI ratings for the videotapes to the "gold standards" that have been developed for each practice video. Essentially, gold standards are "true scores" that have been developed by expert PIs for each videotape used in training

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. Specifically, 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.

Module 5. Performance feedback

The final module is designed to provide PI trainees with feedback regarding the deviation of their practice ratings from the gold standards. The discussion should begin by emphasizing that the purpose of the session is to ensure there are no systematic differences among PI ratings of aircrew performance, explaining how the Gold Standards were developed so the PI trainees view them as "expert" ratings, and explaining the concept of deviation scores and their interpretation.

Once the form and purpose of the feedback are explained, feedback is delivered to the PIs on an item-by-item basis, allowing each PI to compare his or her ratings to that of the experts. For those ratings identified as discrepant from the gold standards a group discussion is facilitated to provide supporting evidence and clarify issues. Gold standard rationales should be used throughout this discussion to help new PIs understand why the experts graded the crew's performance on the video as they did.

The final step involves determining the extent to which PI skills have improved as a result of the training using post-training videotape. PI trainees view the videotape and complete the corresponding grading sheet. These ratings are compared to the gold standards by calculating deviation scores. These deviation scores are then compared to those calculated from initial practice videos. The deviation scores from the post-training exercise provide a measure of skill improvement among the PI trainees. Due to time constraints feedback may be provided on an individual basis at a later time (e.g., via e-mail) to help trainees gauge their individual progress.

Course Length

The time required for gold standards training depends in large part on the number of scenarios and component event sets on which instructors are trained. Our experience suggests that for a single scenario, which consist of three events, gold standards training can be conducted in one day. Modules 1 through 4 can be completed in the morning, and the course instructor can analyze the students practice ratings during the lunch break. The afternoon is devoted to feedback and discussion of the practice ratings (Module 5). For example, one airline carrier has successfully developed and implemented a one-day course using the six modules outlined above. Obviously, more extensive training and greater reliability could be obtained with a longer course.

Tools Needed

A gold standards training program requires a classroom, videotapes, large monitor, VCR, overhead projector, and copies of the scenario grade sheets. A laptop computer may be useful for analyzing performance ratings collected during Module 4. Gold standards training may be delivered via computer-based instruction. Goldsmith and Johnson (in press) have developed a CBI prototype.

References

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

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

Baker, D. P., Swezey, R. W., & Dismukes, R. K. (1998). *A methodology for developing gold standards for rater training videotapes*. Washington, DC: Federal Aviation Administration, Office of the Chief Scientific and Technical Advisor for Human Factors.

Goldsmith, T. E., & Johnson, P. J. (in press). Assessing and improving evaluation of aircrew performance. *International Journal of Aviation Psychology.*

Holt, R. W., Hansberger, J. T., Boehm-Davis, D. A. (in press). Improving rater calibration in aviation: A case study. *International Journal of Aviation Psychology*.

ANNEX

GOLD STANDARDS COURSE DESIGN GUIDE

AGENDA

8:00 - 8:30	**INTRODUCTION**
8:30 - 9:30	**REVIEW PERFORMANCE STANDARDS**
9:30 - 10:00	**BEHAVIORAL OBSERVATION TRAINING**
10:00 – 12:30	**PRACTICE RATING VIDEOTAPES**
12:30 - 13:30	**LUNCH BREAK**
13:30 - 16:00	**PERFORMANCE FEEDBACK**

COURSE DESCRIPTION AND OVERVIEW

Course: Introduction

Instructional Objectives: 1.A through 1.C

Time: 8:00 - 8:30

Description

This module provides new PIs with general background information regarding the role of PIs, 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 PIs;
- 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 PIs. 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 PIs. Describe how Gold Standards will help new PIs adopt a common frame of reference when evaluating crews in the simulator.	C.1
✈ Describe the mechanics of Gold standards training. Emphasize that new PIs 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 instructors.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
1.A.1) Describe the major tasks required of PIs 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 PIs 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 PIs 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: Review Performance Standards

Instructional Objectives: 2.A through 2.E

Time: 8:30 - 9:30

Description

This module provides instruction on the various types of assessments PIs are required to make, the grade sheets they will be using, and relevant performance standards that will be used in the evaluation.

Upon completing this module, trainees will be able to:

- describe the scales used for CRM and TECHNICAL topics, CRM and TECHNICAL 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 TECHNICAL and CRM event set grades;
- describe the process by which TECHNICAL 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 TECHNICAL topic grades, CRM and TECHNICAL event set grades, and PIC and SIC overall grades. CRM and TECHNICAL topic grades refer to broad classes of behavior that can be directly observed. CRM and TECHNICAL 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 crewmember's performance throughout the event set. These grades are based upon the CRM and TECHNICAL topic and event set grades plus the PI's judgment.	A.1
✈ Describe the scale that is used to grade CRM topics. Emphasize that a "Missed observation" means that the PI 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 A.3
✈ Describe the scale that is used to grade TECHNICAL topics. Emphasize that a grade of "1" (Repeat) for a TECHNICAL topic does <u>not</u> <u>require</u> a repeat.	A.4
✈ Describe the scales that are used to grade CRM and TECHNICAL event set performance. Again, emphasize that a grades of "1" (Repeat) do <u>not</u> require a repeat.	A.5
✈ 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 <u>requires</u> a repeat of the event set or parts thereof.	

MAJOR POINTS	**ENABLING OBJECTIVES**
✈ Describe how to grade crew performance on CRM and TECHNICAL 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 TECHNICAL 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: TECHNICAL performance for the event set is graded as "1" if two or more TECHNICAL topics are graded as less than "Standard" <u>or</u> any TECHNICAL 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 TECHNICAL proficiency coupled with the PI's judgment.	C.1
✈ Describe the relative importance of CRM and TECHNICAL 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 PIs 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 TECHNICAL topic grades, CRM and TECHNICAL 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 TECHNICAL topics 2. Overall CRM and TECHNICAL 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 TECHNICAL topics.	Tutorial	Overheads	Oral	TECHNICAL 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 TECHNICAL topic grades, CRM and TECHNICAL 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 TECHNICAL event set performance.	Tutorial	Overheads	Oral	CRM and TECHNICAL event set performance 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
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) <u>requires</u> 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 TECHNICAL 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 TECHNICAL 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 TECHNICAL event set grades are computed.	Tutorial	Overheads	Oral	CRM and TECHNICAL event set grades are calculated using success criteria listed on the grade sheet. There are separate success criteria for CRM and TECHNICAL 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 PI'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 <u>not</u> 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 <u>require</u> 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: Behavioral Observation Training

Instructional Objectives: 3.A through 3.B

Time: 9:30 - 10:00

Description

This module provides instruction on improving new PIs' 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 3.A: To enable trainees to distinguish between behaviors and conclusions.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
3.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 PI'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 3.B: To enable trainees to identify and describe five guidelines for effective observation.

*Enabling Objectives	Strategy	Media	Evaluation	Instructional Content	Type of Learning
3.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: Practice Rating Videotapes

Instructional Objectives: 4.A through 4.B

Time: 10:00 - 12:30

Description

This module provides new PIs 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: Performance Feedback

Instructional Objectives: 5.A - 5.B

Time: 13:30 - 16:00

Description

This module provides feedback that compares the each new PI'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 2 "Performance Standards." PIs 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 PIs using a common frame-of-reference. Emphasize that gold standards training was developed to ensure that all crewmembers would be evaluated consistently, regardless of which PI evaluates them.	A.1
✈ Describe how the Gold Standards represent the ratings of a panel of expert PIs. 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 PI'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 PIs leave training with a common frame-of-reference.	A.4
✈ Have PIs 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

PERFORMANCE FEEDBACK

OBJECTIVE 5.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
5.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
5.A.2) Describe the process by which gold standard ratings were developed.	Tutorial	Overheads	Oral	Groups of expert PIs 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
5.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
5.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
5.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.**

PERFORMANCE FEEDBACK

OBJECTIVE 5.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
5.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. PI 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.**

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 2003	Technical Memorandum

4. TITLE AND SUBTITLE
A Gold Standards Approach to Training Instructors to Evaluate Crew Performance

5. FUNDING NUMBERS
711-31-32

6. AUTHOR(S)
David P. Baker and R. Key Dismukes

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
NASA Ames Research Center
Moffett Field, California 94035-1000

8. PERFORMING ORGANIATION REPORT NUMBER
IH-048

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA/TM—2003–212809

11. SUPPLEMENTARY NOTES
Point of Contact: R. Key Dismukes, M/S 262-4, Ames Research Center, Moffett Field, CA 94035 (650) 604-0150

12A. DISTRIBUTION/AVAILABILITY STATEMENT
Subject Category: 03-06 Distribution: Public
Availability: NASA CASI (301) 621-0390

12B. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The Advanced Qualification Program requires that airlines evaluate crew performance in Line Oriented Simulation. For this evaluation to be meaningful, instructors must observe relevant crew behaviors and evaluate those behaviors consistently and accurately against standards established by the airline. The airline industry has largely settled on an approach in which instructors evaluate crew performance on a series of event sets, using standardized grade sheets on which behaviors specific to event set are listed. Typically, new instructors are given a class in which they learn to use the grade sheets and practice evaluating crew performance observed on videotapes. These classes emphasize reliability, providing detailed instruction and practice in scoring so that all instructors within a given class will give similar scores to similar performance. This approach has value but also has important limitations; (1) ratings within one class of new instructors may differ from those of other classes; (2) ratings may not be driven primarily by the specific behaviors on which the company wanted the crews to be scored; and (3) ratings may not be calibrated to company standards for level of performance skill required. In this paper we provide a method to extend the existing method of training instructors to address these three limitations. We call this method the "gold standards" approach because it uses ratings from the company's most experienced instructors as the basis for training rater accuracy. This approach ties the training to the specific behaviors on which the experienced instructors based their ratings.

14. SUBJECT TERMS
Crew performance evaluation, Line oriented simulation, Gold standards

15. NUMBER OF PAGES
34

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited