THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, *12*(3), 205–222

FORMAL PAPERS

A Framework for Understanding Crew Performance Assessment Issues

David P. Baker
American Institutes for Research
Washington, DC

R. Key Dismukes
National Aeronautics and Space Administration Ames Research Center
Moffett Field, CA

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 issue 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.

The focus of this special issue is on training pilot instructors to assess crew performance. This is a critical matter because few jobs are scrutinized as closely as that of an airline pilot (Goldsmith & Johnson, this issue). All pilots go through extensive qualification training initially and, at a minimum, return on an annual basis for recurrent training.

Requests for reprints should be sent to David P. Baker, American Institutes for Research, 1000 Thomas Jefferson Street NW, Washington, DC 20007. E-mail: dbaker@air.org

Traditionally, airline training and evaluation focused on a pilot's technical skills, which have specified performance parameters (e.g., executing a steep turn with no more than a 100-ft deviation in altitude and a 5-kt deviation in airspeed). The introduction of line-oriented flight training (LOFT) in the 1980s and the Advanced Qualification Program (AQP) in the 1990s launched new requirements for pilot instructors to train and assess not only specific technical flying proficiency, but also complex crew resource management (CRM) skills (Federal Aviation Administration, 1990a, 1990b). LOFT provides aircrews with technical and CRM skills training, whereas the AQP includes technical and CRM skills training and evaluation. Under AQP, formal evaluations occur during line operational evaluations (LOEs), which resemble LOFT except that individual pilot and crew performances are graded. During LOE, a pilot instructor[1] sits in the back of the simulator cab (and typically runs the simulation), observes the crew, and rates the performance of the crew and each crewmember on technical and CRM skills in complex situations, such as understanding and using the airplane's computer correctly to manage the flight through automated systems and exercising good judgment and decision making in ambiguous situations involving weather or equipment malfunctions. The resulting LOE performance ratings serve multiple purposes in the AQP. These include the following: (a) determination of pilots' readiness for line operational duties, (b) assessment of AQP training effectiveness, (c) detection of skill deficiencies among individual pilots, and (d) detection of performance problems across the pilot population (e.g., mode awareness problems with automated aircraft; Birnbach & Longridge, 1993).

Reliable and valid assessment of an aircrew cannot be made during LOE (or any other training and evaluation event) if pilot instructors do not agree on the types of crew behaviors observed and the level of performance these behaviors represent. When pilot instructors do not agree, performance ratings are a function of the particular instructor conducting the assessment as opposed to performance of the crew. If variability among pilot instructors is large, assessment of training effectiveness and crew capabilities is undercut, and the airline may fail to detect performance problems that threaten flight safety.

This special issue presents a set of articles that explore facets of crew-performance assessment and discuss specific strategies for training pilot instructors to accurately assess aircrew performance. In this opening article, we 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 in training pilot instructors and to introduce the articles that comprise this special issue. We also look to literature from domains outside

[1]The term *pilot instructor* is used throughout this article. It encompasses any qualified individual involved in training and evaluating aircrew performance: instructors, check airmen, and standards captains.

of 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 remain to be answered.

A FRAMEWORK FOR UNDERSTANDING CREW-PERFORMANCE ASSESSMENT

In this section we propose a framework for understanding the issues that are encountered in attempting to assess crew performance in a meaningful fashion. We do this by first examining a specific case of aircrew evaluation—LOE—and then derive our framework from this process. Although we specifically focus on LOE, we believe that facets of this framework are generalizable to any situation in which a pilot instructor observes and assesses aircrew performance.

Line Operational Evaluation

The framework for LOE was initially begun through innovative collaborative work by individuals in several organizations (Hamman, Seamster, Smith, & Lafaro, 1991). In current practice, LOE varies among airlines in some details, but the framework generally follows the original design. LOE scenarios are designed around a series of event sets in which aircrews encounter situations designed to test their CRM and technical skills. Typically, a specific phase of flight (e.g., taxi or takeoff) or a combination of several phases of flight defines an event set. Usually a trigger, for example, the failure of the nose gear to retract during takeoff, initiates an event set. Using a grade sheet specifically designed for each event (Figure 1), a pilot instructor observes and evaluates an aircrew's response to the event set (i.e., how they handle the nose-gear problem). Pilot instructors also grade crew performance across event sets on the entire LOE scenario.

Research has shown that the event-set methodology produces more reliable performance ratings than holistic judgment strategies (Seamster, Edens, & Holt, 1995). However, this process places considerable demand on pilot instructors because of the number of event sets that must be rated and the number of ratings required for each event. These demands are one source of variability in ratings among instructors. Another source of variability lies in the structure and corresponding requirements of the grade sheet that is used by pilot instructors to record observations and rate crew performance on the LOE.

Figure 1 depicts an example of a grade sheet for an event set that is triggered by an oil filter Engine Indication and Crew Alert System (EICAS) message. The grade sheet lists several CRM behaviors that crew members are expected to exhibit. Pilot instructors rate each CRM behavior as *not observed, partially*

EVENT SET TRIGGER: Left or Right engine "OIL FILTER" EICAS message				
CRM Behaviors: Provide a comments when marking "Not Performed" for clarification	Missed	Not Performed	Partially Performed	Performed
a. Task assignment and prioritization	O	O	O	O
b. Open communication and crew member respect	O	O	O	O
c. Information gathering for situation analysis	O	O	O	O
d. Deciding on a plan and the plan implementation	O	O	O	O
TECHNICAL Behaviors: Comments are required for a "1" or "2" and requested for a "4"	Repeat	Debriefed	Standard	Excellent
e. Company communications	O	O	O	O
f. Engine irregular/Emergency procedures	O	O	O	O
g. Diversion operations	O	O	O	O
EVENT SET GRADE: Select only one grade category	Repeat	Debriefed	Standard	Excellent
CREW TECHNICAL	O	O	O	O
CREW CRM	O	O	O	O
COMMENTS: Provide written description of observations in this event set. The description must identify the event set topics by letter marked above. Event set grades of repeat or debriefed require comments. Comments are requested for excellent.				
OVERALL GRADE	Repeat	Debriefed	Standard	Excellent
Captain/Overall	O	O	O	O
First Officer/Overall	O	O	O	O

FIGURE 1 Example grade sheet.

observed, or *observed* during the event set. *Observed* means the instructor saw the desired behavior; *partially observed* means the instructor saw some but not all aspects of the desired behavior. In addition, a rating of *missed* can be recorded if the instructor believes that he or she missed the observation because of other tasks he or she must perform while conducting the LOE (e.g., running the simulator or role playing air traffic control communications). Technical behaviors are graded on a different scale: *repeat, debriefed, standard,* and *excellent. Repeat* means performance on the event set must be repeated after additional training. *Debrief* means some inadequacies were observed and must be discussed. This

scale is also used to assess each crew's overall CRM and technical performance on the event set as well as each individual pilot's performance.

In theory, observations of specific CRM behaviors and ratings of specific technical behaviors serve as the basis for rating overall CRM performance and technical performance as well as rating each individual crew member's performance on the event set. (However, research presented in this special issue raises questions about whether instructors actually base their overall ratings primarily on the specific observable behaviors identified by the grade sheet. See Holt, Hansberger, & Boehm-Davis, this issue, and O'Connor, Hörmann, Flin, Lodge, Goeters, & the JARTEL Group, this issue.) Performance across event sets is supposed to serve as the basis for grading pilot and copilot performance on the overall LOE scenario. This process of observation, evaluation, and judgment by the pilot instructor is depicted in Figure 2.

Components of the Framework

In our framework, the assessment of crew performance requires three critical activities: (a) observation of specific technical and CRM behaviors, (b) evaluation of these behaviors with respect to their effectiveness, and (c) weighing the results of this behavior evaluation process to arrive at different scores that must be recorded on the grade sheet. Instructors may differ in how they perform each of these three activities, leading to differences among instructors' grades for a given aircrew in a given event set (Borman, 1978). For example, different pilot instructors may attend to different aspects of behavior in an event set. Grade sheets typically list only a small number of behaviors to be observed and assessed from a much larger set of behaviors that the crew is engaged in during an event set. Limiting the number of behaviors to be rated may seem to simplify the instructor's task, but unfortunately it is complicated by the fact that the crew's behaviors are highly interrelated, thus making it difficult to consider just one aspect in isolation from the rest. In addition, the behaviors listed on the grade sheet often vary in specificity of description. Thus, although grade sheets appear to be quite explicit and specific in what behaviors are to be rated, the structure actually leaves considerable room for individual differences among instructors to operate.

Personal construct theory suggests that in these sorts of situations, individuals are likely to use unique personal constructs to evaluate information (Bannister & Mair, 1968; Kelly, 1955). What is important to one pilot instructor may be less important to another instructor. Also, instructors may use different strategies to combine evaluations of specific crew behaviors to arrive at event-set and scenario grades for the crew (i.e., CRM performance and technical performance) and each pilot. For example, in reference to Figure 1, some pilot instructors might view the crew's strategy for assigning and prioritizing tasks in relation to the EICAS

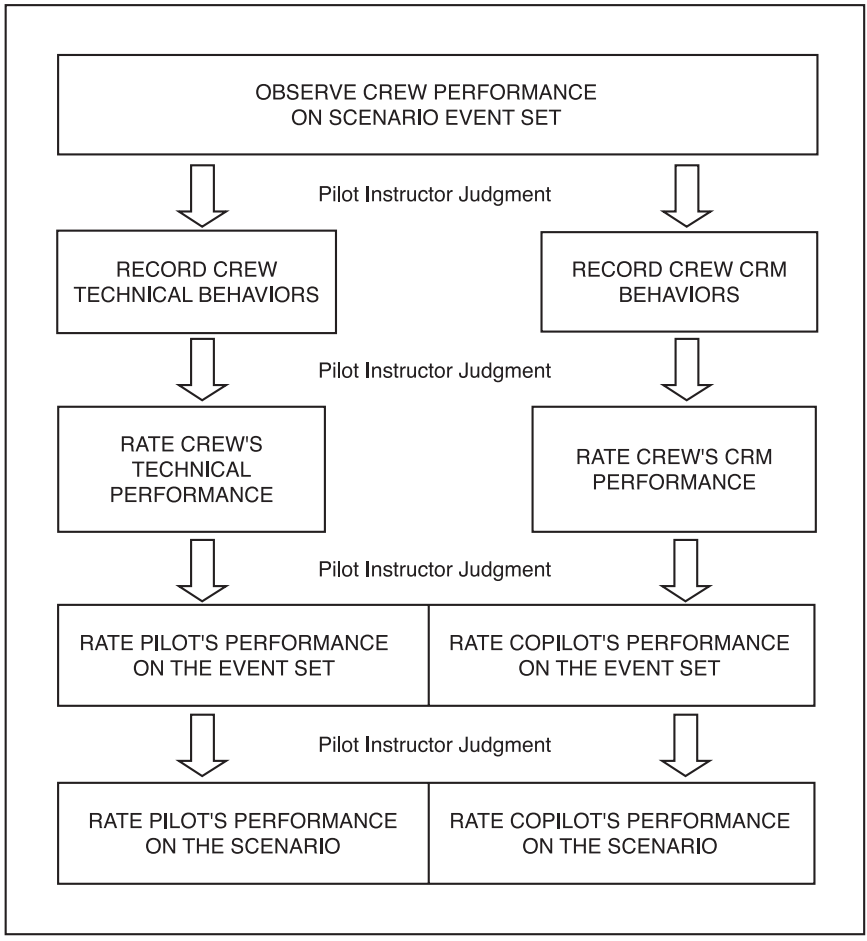

FIGURE 2 A framework for understanding crew-performance assessment.

message as more important than how the crew decides on a plan and implements it. Assigning different weights to the four CRM behaviors listed on the grade sheet will result in different event-set grades for CRM for different pilot instructors.

In addition to the sources of inaccuracy described previously, Borman (1978) cited three additional problems that are relevant to crew-performance assessment. First, pilot instructors may not have sufficient opportunity to observe relevant CRM and technical behaviors during a scenario. Brannick and his colleagues (Brannick, Prince, & Salas, this issue; Brannick, Roach, & Salas, 1993) argue that some scenarios, as currently constructed, do not elicit enough behavioral information to produce reliable judgments on the part of pilot instructors. Also,

instructors are typically very busy running the simulation and cannot devote full attention to observing and evaluating the crew.

Second, real or implied organizational constraints (within the fleet or the entire airline) may distort the crew assessment process. For example, pilot instructors may be implicitly encouraged to reduce the range of possible ratings on the grade sheet (e.g., *excellent, standard, debrief,* and *repeat*) to a scale that reflects acceptable or unacceptable performance. Our informal review of several air-carrier grade sheets indicates that extreme grades (positive or negative) often require written justification by pilot instructors, but grades using the middle ranges of the grading scale do not. The extra effort required for writing these justifications may unconsciously bias the instructors against using the extreme ratings, although they would not deliberately give a passing grade to an unsafe pilot. Reducing the range of grades used throws away valuable information the airline needs to evaluate training effectiveness and performance trends.

Finally, pilot instructors bring different levels of experience to the LOE evaluation task. Experience may vary as a function of the LOE scenario, particular event sets, or the grade sheets used. Also, most instructors have not been systematically trained to observe behaviors, weigh and integrate observations, and assign grades. Thus the airlines must provide explicit training and practice in these processes to avoid substantial variability among instructor ratings.

Implications for Training Pilot Instructors

The LOE scenario event-set methodology and the introduction of highly structured grade sheets for assessing crew performance on each event set have led to improvements in pilot instructor accuracy (O'Connor et al., this issue; Seamster et al., 1995). However, in other domains, these strategies have been shown not to produce adequate rater accuracy unless supplemented with explicit rater training (Landy & Farr, 1980). Moreover, under the AQP, pilot instructor training is required and instructor calibration must be assessed periodically (Birnbach & Longridge, 1993).

The framework presented in Figure 2 has several implications for training pilot instructors to accurately assess crew performance. Our focus is on accuracy, because accurate ratings lead to high levels of instructor agreement and reliability, but high levels of agreement and reliability do not necessarily result in accurate ratings (see Goldsmith & Johnson, this issue, for a detailed discussion of these concepts and the relationships among them). Figure 2 illustrates three critical areas in which pilot instructors should be trained:

1. Accurate observation of crew behavior during realistic full-mission scenarios. Observations serve as the basis for performance ratings, so accurately observing and recording crew behaviors during each event set is critical.
2. Evaluation of behavioral information. Criteria for judging the effectiveness and ineffectiveness of different crew behaviors should be provided so

instructors use the same standards when judging crew performance. Criteria must be explicit and clearly applicable to the event sets to be evaluated.

3. How to assign crew technical and CRM performance ratings and how to rate each individual pilot's performance on the LOE. Assigning performance grades involves weighing and combining behavioral judgments to arrive at different performance scores (Borman, 1978).

Currently, airlines have little to guide them in how best to provide instructors with training in these three areas. The articles in this special issue provide some initial insight into each of these issues. These articles reflect the current state of the research on training pilot instructors to assess crew performance. We briefly review each of these articles and discuss how they contribute to the three training needs outlined previously. In the section that follows we review literature outside aviation psychology that suggests strategies for training instructors to assess crew performance accurately.

SPECIAL ISSUE ARTICLES

In addition to this opening article, five articles are presented in this special issue that focus on training pilot instructors to assess crew performance. Four of these articles present empirical data (Brannick et al., this issue; Holt et al., this issue; Mulqueen, Baker, & Dismukes, this issue; O'Connor et al., this issue), whereas one is conceptual in nature (Goldsmith & Johnson, this issue). Each article addresses different components of the framework presented in Figure 2.

Goldsmith and Johnson (this issue) provide an informative discussion of the application of statistical methods for assessing and improving the quality of evaluation of aircrew performance. Specially, they describe measures of interrater and referent reliability and the application of these methods for training pilot instructors to assess crew performance. In addition, Goldsmith and Johnson provide an insightful discussion of the relationships between rating reliability and accuracy and provide a strong argument on why pilot instructor rater training should target rater accuracy as opposed to interrater reliability (IRR). Finally, these authors describe an approach for training pilot instructors that they have implemented at several air carriers.

Brannick et al. (this issue) present reliability data for two instructors who rated the performance of 45 helicopter crews who flew a simulated mission in a full-motion simulator. These instructors were trained with a combination of performance dimension training (PDT) and behavior observation training (BOT) before rating the helicopter crews. A detailed description of this training is provided and IRR and internal consistency estimates are presented for ratings of specific crew behaviors, each scenario event, and several CRM skills for the entire scenario

(e.g., decision making). Although a direct examination of the effectiveness of PDT and BOT was not conducted, this study presents information on the levels of reliability that can be achieved when two instructors are trained under ideal conditions.

O'Connor et al. (this issue) outline the development of a European behavioral marker system for CRM evaluation called NOTECHS. Although this article only indirectly addresses pilot instructor rater training, significant insights are provided regarding the development of a generic grade sheet for crew-performance assessment. A preliminary test of NOTECHS was conducted with 105 pilot instructors from 14 European airlines. After the training phase, these instructor pilots used the system to evaluate the individual CRM skills of captains and first officers in eight different video scenarios filmed in a Boeing 757 simulator. Information on the reliability and accuracy of the NOTECHS system is presented.

Mulqueen et al. (this issue) use a multifaceted one-parameter item response theory (i.e., Rasch) model to analyze the quality of a pilot instructor rater-training program. The Rasch model provides a means for examining individual pilot instructor leniency or severity in ratings, difficulty of grade-sheet items, skill levels of flight crews, and interactions among these components. Therefore it provides a comprehensive examination of a rater-training program, because it provides information on the quality of pilot instructor ratings, the quality of LOE grade sheets, and the quality of the videotapes used in training.

Holt et al. (this issue) describe a case study at a regional air carrier focused on improving the reliability and validity of crew-performance assessments. These researchers evaluated their approach for training pilot instructors to assess crew performance in a 3-year study at a regional airline. They constructed and evaluated five metrics for assessing IRR and designed a standardized process for using these metrics to train pilot instructors. Data on the quality of crew-performance assessment are reported for two fleets at the air carrier. Holt et al. also provide an interesting examination of the structural validity of the crew-performance assessment process that is represented in Figure 2.

In summary, the articles presented in this special issue either directly or indirectly address different facets of the framework presented in Figure 2. First, in regard to the accurate observation of crew behavior, the Brannick et al. (this issue), Holt et al. (this issue), and O'Connor et al. (this issue) articles all provide strategies for enhancing the accurate observation of crew behavior and data on the levels of reliability and accuracy that can be achieved for behavioral observation. The strategies for improving observation in these articles primarily involve modification of grade sheets, but recommendations are also made on how to train this aspect of the crew-performance assessment process. Second, to improve evaluation of crew behaviors, Goldsmith and Johnson (this issue), Holt et al. (this issue), and Mulqueen et al. (this issue) all present strategies for providing pilot instructors with feedback during training about weighting and evaluating behavioral information.

These procedures tend to rely on analysis of crew-performance ratings and discussion of discrepancies among pilot instructors to establish standards across instructors at the airline. Finally, regarding the rating of crew performance, all of the articles provide insights in different ways. Some recommend strategies for training (e.g., Brannick et al.; Goldsmith & Johnson; Holt et al.), some recommend strategies for data analysis and feedback (Holt et al.; Mulqueen et al.), and some provide insights on grade-sheet design (Brannick et al.; O'Connor et al.).

RATER-TRAINING STRATEGIES

Performance Appraisal

A substantial amount of research on the effectiveness of different strategies for training raters has previously been conducted in the domain of performance appraisal in which a supervisor observes and evaluates subordinate performance on various job-related dimensions (Borman, 1978; Smith, 1986; Woehr & Feldman, 1993; Woehr & Huffcutt, 1994). This process is similar to crew-performance assessment except that the evaluation of crew performance involves observing and assessing aircrews flying simulators rather than employees performing their jobs. We believe that much of what is known about training supervisors to conduct accurate performance appraisals is directly generalizable to aircrew evaluation and should be leveraged to develop pilot instructor training.

Overall, there are two well-established findings in the literature regarding how to develop and deliver effective rater training. First, some strategies are more effective than others are. Rater error training, PDT, BOT, and frame-of-reference (FOR) training are among the strategies that have been examined in detail. Results from a detailed meta-analysis found that FOR training is the most effective strategy for improving rating accuracy, whereas BOT is most effective for improving observation accuracy (for a comprehensive review, see Woehr & Huffcutt, 1994). Second, significant gains in accuracy occur when rater training includes opportunities for trainees to practice and receive feedback on the rating task (Smith, 1986). BOT, FOR, and the nature in which these strategies differ from current pilot instructor rater training are briefly discussed next.

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, whereas evaluation processes include categorizing, integrating, and evaluating information. In BOT, faulty behavioral observation is viewed as the primary reason for rating inaccuracies. Typically BOT encompasses strategies that focus on observation or recording of behavior (e.g., note taking and diary keeping). Discussion

and practice exercises that focus on recognizing and avoiding systematic errors of observation, contamination from prior information, and overreliance on a single source of information may also be included (Thornton & Zorich, 1980; Woehr & Huffcutt, 1994).

The purpose of FOR training, as the name implies, is to train raters to a common frame of reference. FOR training, in many aspects, resembles IRR training, which is currently used by several airlines to train pilot instructors under the AQP (Holt et al., this issue; Mulqueen et al., this issue). Pilot instructors are presented with information about the LOE scenario, the grade sheet, and the relevant technical and CRM skills to be assessed. They are also given practice and feedback on the LOE assessment task. Practice usually involves rating a series of videotapes of different aircrews flying several LOE scenario event sets, and feedback usually involves instructors comparing their ratings and resolving differences through discussion (Goldsmith & Johnson, this issue; Holt et al., this issue).

However, FOR training differs from IRR training in one crucial aspect. Feedback in FOR training compares each pilot instructor's ratings with a set of previously defined true scores. True scores have been referred to as gold standards in the air-carrier industry. Gold standards are assigned to each crew depicted on a videotape by expert pilot instructors who watch the videotape, independently assign CRM and technical performance ratings on the LOE grade sheet, and discuss their ratings to reach consensus. Gold standards may also include a description of the behaviors that drive the scores. The resulting ratings are taken to be the best assessment obtainable of actual crew performance on the scenario event sets.

Gold-Standards Training

Empirical research from the field of performance appraisal demonstrates the potential effectiveness of BOT and FOR training for training pilot instructors to accurately assess aircrew performance during LOE. However, these strategies have yet to be applied at an air carrier, making their structure and application somewhat unclear. In this section, we describe what a combination of BOT and FOR training might look like at an airline. We refer to this training as gold-standards training, because it combines the most desirable characteristics of BOT, FOR, and IRR and relies on gold standards for providing pilot instructors with feedback about their rating accuracy. Currently, we are developing this training at a major U.S. air carrier. Once developed, this program will be implemented and tested at the carrier.

Gold-standards training should initially include a detailed review of the LOE scenarios to be evaluated. Although this is not a defining feature of BOT or FOR training, this practice is included in IRR and is important for developing a common frame of reference among instructors at the airline. In cases in which pilot

instructors are being trained for the first time, this review should also include a detailed explanation of the LOE grade sheets, including the rating scales and grading rules (e.g., cases in which certain behavioral observations lead to specific performance ratings). In cases in which pilot instructors are receiving recurrent training, any changes to the grade sheet or the grading process should be noted and discussed.

In addition to reviewing the LOE scenario and the grade sheets, gold-standards training should include a review of the performance standards for each technical and CRM skill to be assessed. Information regarding technical and CRM requirements for successful performance on each LOE event set is often found in the airline's qualification standards. Similar to behaviorally anchored rating scales, this information can be used to develop specific examples of different levels of performance on the grade sheet. For most air carriers, examples of excellent, standard, debriefed, and repeat levels of performance should be developed for each event set. Pilot instructors could then use these examples as referents during the LOE rating process, which should enhance rating accuracy.

To ensure observation accuracy, BOT should be included in gold-standards training. 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, and verifiable) and how to accurately observe an aircrew's performance during LOE. Discussion may be particularly beneficial in recurrent gold-standards training, because pilot instructors could share their experiences regarding observation strategies that they have found to be effective and ineffective. Second, observation training should include opportunities for practice and feedback. The research on rater training indicates that practice and feedback are critical for training transfer (Smith, 1986). Therefore, pilot instructors should be shown a series of videotapes for the purpose of practicing their observational skills. The videotapes should be annotated with detailed observations from experts about the specific behaviors exhibited by the crews 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 with observations and interpretations of experts.

Finally, gold-standards training should include practice and feedback with the rating task. Ideally, this practice should include rating the videotaped performance of crews flying event sets from the LOE scenario that will be rated by pilot instructors in the future. Furthermore, gold-standards training should consist of practice videos that display a range of crew-performance levels. Here, we recommend including a minimum of at least three practice videotapes displaying excellent, average, and poor crew performance. However, the specific number and types of practice tapes that should be included to ensure the highest probability of training transfer has yet to be determined empirically.

Because of the strong empirical support for FOR training, gold-standards training should include feedback based on gold standards that are developed by expert pilot instructors. In addition, feedback should include information on expert rationales for each gold standard. Specific methodologies for developing gold standards have been presented in the literature (Baker, Swezey, & Dismukes, 1998; Bernardin & Buckley, 1981; Goldsmith & Johnson, this issue). The research demonstrates the importance of gold standards for training pilot instructors to rate aircrew performance the way expert instructors perform ratings. Furthermore, by using the same gold standards across pilot instructor training classes to provide feedback, as opposed to norming instructors to the standards established in each training class (i.e., as performed in IRR training), greater reliability and accuracy should be observed across classes. Overall, this approach should improve the quality of crew performance evaluations made during LOE.

DETERMINING TRAINING EFFECTIVENESS

Our literature review suggests that gold-standards training should improve pilot instructor accuracy when assessing crew performance during LOE, and we have presented an approach for implementing such training. However, research has yet to empirically establish the effectiveness of this approach in the airline industry. Such research is critical to ensure that (a) training is effective, (b) correct decisions are made about pilot certification, and (c) useful and informative data are collected about AQP effectiveness. To help guide training effectiveness research, Sackett and Mullen (1993) delineated two specific questions that training effectiveness studies can seek to answer: How much change has occurred? What level of performance has been achieved? Each of these issues is important in the context of training pilot instructors to assess crew performance for different reasons, which are briefly discussed next.

Sackett and Mullen (1993) outlined three situations that call for research that assesses change measurement: (a) when one wishes to determine the utility of a training program, (b) when one wishes to compare the efficacy of different training programs, and (c) when one wishes to make a contribution to training research. Change measurement involves using a proper experimental design to test a specific research question. Here, the focus is on experimental rigor (e.g., Was a control group used? Was there random assignment to training conditions?), statistical significance (e.g., Did one training program produce pilot instructors who were significantly more accurate at assessing crew performance?) and meaningfulness of results (e.g., How large was the effect?).

Change measurement research is most important in the current context from the standpoint of identifying the most effective strategies for training pilot instructors to assess crew performance. Brannick et al. (this issue), Holt et al.

(this issue), and O'Connor et al. (this issue) all report positive results for their instructor training, but none of these studies directly compares the efficacy of different training strategies. Therefore, as part of this special issue, we call on aviation researchers to conduct such empirical tests. In particular, we advocate that studies examine the utility of different strategies for training pilot instructors to assess crew performance. Is IRR the best method for training pilot instructors or is gold-standards training or some other strategy a better alternative? Proper experimental or quasi-experimental design should be used to answer such questions and to ensure the integrity of results on which future decisions about pilot instructor rater training will be based.

In addition to determining how much change has occurred, training effectiveness studies should examine whether or not a specific level of performance has been achieved. Sackett and Mullen (1993) suggested that this research question is appropriate when a clear target level of performance exists and an organization is interested in documenting the performance of specific employees. Determining whether a specific level of performance has been achieved is fairly simple if a criterion for adequate performance exists; trainee performance can be measured against this criterion to establish training effectiveness.

Determining whether or not pilot instructors have achieved a specific level of performance is also important in the current context, because airlines should know how accurate pilot instructors are at assessing crew performance and whether or not pilot instructors as a group are in agreement (i.e., interchangeable). Simply knowing which training strategy is most effective is not enough; research should determine what level of accuracy and reliability are produced by different training strategies. Goldsmith and Johnson (this issue), Holt et al. (this issue), and Mulqueen et al. (this issue) present some specific strategies for assessing pilot instructor accuracy and reliability. Holt et al. also provide some initial data on the levels of agreement, consistency, sensitivity, and congruency that can be achieved for pilot instructors at a regional air carrier. These data could be used as the starting point for developing performance standards for pilot instructors. If acceptable standards could be developed, airlines could use these standards for certifying pilot instructors as evaluators. For example, only instructors who deviate a certain number of scale points from the expert gold standard might be certified to evaluate crew performance on an LOE. Instructors not meeting this criterion could be offered additional training until they achieve a desired level of performance established by the airline.

FUTURE RESEARCH

Although establishing the most effective strategy for training pilot instructors to assess crew performance is paramount, as well as establishing levels of perfor-

mance that can be expected from such training, numerous research questions will still remain regarding training pilot instructors to assess aircrew performance. Dismukes (1999) outlined many of these issues and Goldsmith and Johnson (this issue) provide a similar discussion. Here we organize these questions into research themes.

Training Effectiveness

What is the most effective strategy for training pilot instructors to assess crew performance? The articles presented in this special issue make some inroads into this issue; however, a direct comparison of different rater-training strategies is required to determine the strategy or strategies that produce the greatest gains in pilot instructor observational and rating accuracy. The literature reviewed suggests that gold-standards training would be an effective approach, but well-designed research is required to test this hypothesis. In addition, training strategies should be compared across multiple pilot instructor rater-training classes. Although a particular training strategy may produce significant gains in observation and rating accuracy within a training class, similar gains in accuracy may not be realized across different pilot instructor training classes (Baker, Mulqueen, & Dismukes, 2001). This issue is important for large air carriers in which large cadres of pilot instructors cannot all be trained at the same time.

Recurrency Training and Transfer Issues

Once the most effective rater-training strategy has been established, a number of other questions should and can be answered. Specifically, research needs to determine at what rate do observational and rating accuracy decay, and how often should instructors be retrained? Should recurrent training be conducted in the same fashion as initial rater training (e.g., formal gold-standards training), or is some other strategy appropriate? Research should also establish the extent to which pilot instructor rater training transfers from the LOE event sets used for practice and feedback during training to similar event sets. At most airlines, pilot instructor training allows practice and feedback on only a limited number of event sets, and these event sets may or may not be part of the LOEs graded by instructors on the line. Are instructors trained on one LOE or one set of the event sets accurate when grading different but similar LOEs or event sets? Research should also establish whether pilot instructor rater training conducted in the classroom transfers to the simulator, which places additional demands on the pilot instructor (e.g., running the simulator and role playing air traffic control). Finally, research should determine if pilot instructor rater training produces pilot instructors who are accurate across a wide variety of aircrews who display a full

range of performance on the LOE event sets. At the heart of this issue is whether some aircrews or some levels of performance are more difficult to evaluate than others.

Structural Validity

Follow-up is required for initial investigations of the structural validity of the crew-performance assessment process. Most LOE grade sheets that we have seen require instructors to actually record their behavioral observations and then use these observations when making crew performance ratings. Under the AQP, observations listed on the LOE grade sheet are often tied to qualification standards that have been established at the airline. Therefore, behavioral observations should account for the majority of variation in the event set and individual crewmember performance ratings. However, results presented by Holt et al. (this issue) indicate that this in fact may not be the case. Possible explanations for this finding include limitations associated with LOE grade sheets (not all important behaviors are listed), individual differences associated with the pilot instructors (pilot instructors weight behavioral information differently as a function of their own personal constructs), and ineffective training (training does not teach pilot instructors to properly use the grade sheets). In addition, Brannick et al. (this issue) show that instructors demonstrate different levels of IRR and internal consistency depending on the types of crew performance rating made. The Brannick et al., Holt et al., and O'Connor et al. (this issue) articles all provide insight regarding the structural validity of the crew-performance assessment process and make recommendations for addressing this issue. However, these recommendations remain to be tested. We believe that this is a critical area to explore because the structural validity of LOE directly contributes to the effectiveness of the AQP. Aircrew performance deficiencies identified during LOE are the basis for adjustment to AQP training. Deficiencies are identified by reviewing information collected on the LOE grade sheets. If this information is unreliable, training developers may make incorrect decisions on how to adjust the airline's AQP training program. Therefore, we call on aviation researchers to continue to explore the structural validity of the LOE grading process and caution airlines to examine the quality of LOE performance data when considering adjustments to their training programs.

SUMMARY

The focus of this special issue is on training pilot instructors to assess crew performance. In this opening article we have introduced a framework for understanding crew-performance assessment. The five articles that follow provide

insight about different facets of this framework. These articles reflect what the research community currently knows about training pilot instructors to assess crew performance. We would encourage readers of this special issue to review each of these articles in light of the framework presented in Figure 2 and to identify areas in which future research can contribute to modifying and validating this framework. Considerable insight can be gained regarding potential training strategies, variants in grade-sheet design, and strategies for analyzing and providing feedback about pilot instructor rating accuracy. Finally, we encourage readers of this special issue to carefully consider the implications of the articles for the AQP and aircrew training in general. We hope that readers of this special issue will recognize the limitations of these studies and that future research will improve the overall quality of the crew-performance assessment process. The articles presented in this special issue are a starting point; however, there is still much to be learned and much work to be performed in regard to training pilot instructors to assess crew performance.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Aeronautics and Space Administration (NASA) Ames Research Center (Grant No. NCC2–1033). The views presented in this article are those of the authors and should not be construed as an official NASA position, policy, or decision unless so designated by other official document.

REFERENCES

Baker, D. P., Mulqueen, C., & Dismukes, R. K. (2001). Within-group versus between-group consistency: Examining the effectiveness of IRR training. *Proceedings of the International Symposium on Aviation Psychology.*

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.

Bannister, D., & Mair, J. M. M. (1968). *The evaluation of personal constructs.* New York: Academic.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205–212.

Birnbach, R. A., & Longridge, T. M. (1993). The regulatory perspective. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 263–281). New York: Academic.

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology, 63,* 135–144.

Brannick, M. T., Roach, R. M., & Salas, E. (1993). Understanding team performance: A multimethod study. *Human Performance, 6,* 287–308.

Dismukes, R. K. (1999). Discussion: Issues in evaluating crew performance in line oriented evaluation. *Proceedings of the International Symposium on Aviation Psychology, 1,* 329–331.

Federal Aviation Administration. (1990a). *Advanced Qualification Program* (Advisory Circular 120–54). Washington, DC: Department of Transportation.

Federal Aviation Administration. (1990b). *Special Federal Aviation Regulation 58—Advanced Qualification Program* (Federal Register, Vol. 55, No. 91, Rules and Regulations, pp. 40262–40278). Washington, DC: National Archives and Records Administration.

Hamman, W. R., Seamster, T. L., Smith, K. M., & Lofaro, R. J. (1991). The future of LOFT scenario design and validation. *Proceedings of the 6th International Symposium on Aviation Psychology,* 589–594.

Kelly, G. A. (1955). *The psychology of personal constructs* (Vol. 2). New York: Norton.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72–107.

Sackett, P. R., & Mullen, E. J. (1993). Beyond formal experimental design: Toward an expanded view of the training process. *Personnel Psychology, 46,* 613–627.

Seamster, T. L., Edens, E. S., & Holt, R. W. (1995). Scenario event sets and the reliability of CRM assessment. *Proceedings of the 8th International Symposium on Aviation Psychology,* 613–618.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Journal, 11,* 22–40.

Thornton, G. C., & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology, 65,* 351–354.

Woehr, D. J., & Feldman, J. M. (1993). Processing objective and question order effects on the causal relation between memory and judgment in performance appraisal: The tip of the iceberg. *Journal of Applied Psychology, 78,* 232–241.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67,* 189–205.

Manuscript first received June 2001