A Guide to the Evaluation Aspects of Entering AQP

Peder J. Johnson and Timothy E. Goldsmith

University of New Mexico
Sponsored by FAA Grant: 94-G-013
Contract Monitor: Dr. Eleana Edens

Over the past two years we have worked with several carriers that are at different phases of development in their AQP program. Our work with these carriers has primarily been related to the development of software to facilitate data collection, statistical analysis, and report generation while conducting Line Oriented Evaluations (LOE) calibration sessions with Instructor/Evaluators (I/Es). During this time we have also assisted the carriers in the design of gradesheets, and the development of a Performance Audit Databases (PADB) and Proficiency Performance Databases (PPDB). As a result of these experiences it has become apparent that there is a set of relatively basic questions and problems that arise when a carrier moves into AQP. In this report we discuss some of the most common of these issues and wherever possible make some suggestions based on our experiences with other carriers.

Obviously, every carrier is somewhat unique in terms of the specific issues that it will encounter. As a consequence, the set of issues we discuss below are best viewed as simply an impetus to stimulate thinking and planning. Finally, although the information in this document is consistent with FAA guidelines for developing and administering an AQP, the carrier should consult SFAR 58 and AC 120-54, as revised, for FAA approved AQP procedures.

Section I: Overview

Goal of AQP: Assurance of Quality and Effective Training

It is essential that all the major participants within a carrier clearly understand that the primary goal of AQP is to ensure that crew training is maximally efficient and effective. Toward this end it is necessary to evaluate crews/pilots to determine whether the training is accomplishing its desired ends, and if not, what specific aspects of training need to be improved. What essentially is being advocated is proficiency-based training and assessment. For this type of a program to work, it is absolutely necessary that the assessment tools are reliable and valid. Thus, AQP directs attention toward the assessment process (i.e., the evaluation of Instructor/Evaluators (I/Es), and the methods and instruments they use). With this increased focus on the assessment of I/Es it becomes easy to lose the forest from the trees. Supervisors and I/Es need to be reminded that assessment is not exclusively about evaluation. Rather, assessment is better viewed as a means to an end-the effective and efficient training of pilots and crews.

Viewing assessment primarily in terms of evaluation can adversely affect the design of assessment instruments. To illustrate, the gradesheets used by many carriers require I/Es to grade concrete or discrete items (e.g., observed behaviors (OBs) or performance indicators) within each event set. However, some I/Es do not see a need for grading pilot performance at this more specific level. Getting down to the level of grading specific concrete behaviors within an event set may be seen as missing the "big picture". Although there maybe some validity to this criticism, it is often made from the perspective of seeing grading as serving only an evaluative function, whereas AQP requires that we also be concerned with the training implications of grading. Event set grades are a good indicator of overall performance, however by themselves they are not sufficiently diagnostic (i.e., they do not point to the underlying

problem). The analysis of OB data aggregated across an entire fleet is far more likely to reveal precisely where training and the curriculum is and is not working.

Primacy of Quality Performance Assessment Data

AQP only works as well as the assessment process is working. If we are unable to reliably assess pilot/crew performance we will not know if our training is effective and it is impossible to determine whether we are training to proficiency. The implications of this are quite obvious. The development of curriculum and training must go hand in hand with the development of the assessment process. We must continually demonstrate that we are collecting quality data. This is one of the reasons why it is so important to calibrate I/Es and the associated assessment tools (e.g., gradesheets). With any proficiency based program, it becomes essential that we objectively determine how reliably we are evaluating pilot performance. We will discuss this is greater detail later when we turn to I/E calibration sessions.

Return to Top

Section II: Developing AQP Databases

The Development of a Program Audit Database (PADB)

Entering AQP requires the carrier to demonstrate that it is training all of the important pilot/crew knowledges and skills to proficiency. To do this we must first know precisely what the proficiencies are. This is accomplished with a front-end analysis of pilot/crew tasks. The overall goal here is to define as precisely as possible what tasks a pilot/crew must be able to do at each phase of flight, what knowledges and skills underlie successful performance of these tasks, and finally what curriculum structure is required to support these knowledges and skills. The idea of a PADB is to explicitly capture all of the above information in the context of a relational database. Specifically, it represents the results of the front-end task analyses that have been performed, including the task list, qualification standards, and curriculum outline. In addition, it provides audit trails that can be used to link curriculum elements to the task list and qualification standards. These audit trails are an important means for ensuring that all aspects of the job (as defined by the task list) and the corresponding qualification standards (which define what constitutes task proficiency) are trained and evaluated.

The Development of a Performance Proficiency Database (PPDB)

AQP also requires that each carrier establish a PPDB which stores de-identified performance data in a manner that allows the carrier to continuously monitor pilot proficiency by conducting statistical analyses on aggregated data. For example, the PPDB could be used to generate a report that would show the average event set grade on Line Oriented Evaluations (LOE) plotted over the past 12 months for a given fleet. The PPDB must address three criteria. First, it must be able to store all relevant performance data that pertain to a fleet's proficiency (e.g., LOE performance, First Look and Maneuvers performance, and Line-check data (Ref: AC1250-15, as revised)). The reports that are required by the FAA as part of AQP should be viewed as setting the database's minimum capability. Over and above the FAA requirements, various questions will arise regarding crew proficiency and the effectiveness and efficiency of training. Ideally, the database would be designed with the foresight that any conceivable question could be addressed by the existing database. While, it is practically impossible to foresee all the questions that may arise in the future, we would recommend that the carrier carefully solicit input from all potential users of the database while it is being developed. To

facilitate this process we (Johnson & Goldsmith, 1998) have generated a list of questions that have arisen in discussions with other carriers when designing a PPDB that is available to all carriers.

In addition to storing all the relevant data, the PPDB must be able to aggregate various subsets of the performance data to address specific questions (e.g., analyze pilot LOE event set performance by I/E) and generate statistical reports (e.g., graphs, histograms, etc.) that are easily interpreted by fleet supervisors. Some of these statistical reports will be generated on regular intervals, whereas others will be ad hoc (e.g., when a regular report suggests a potential problem, the fleet supervisor may what to look more deeply into the data in an attempt to find what is producing the problem). Some of the reports may be designed to provide specific information required by the FAA, whereas most of the reports will quite likely be used internally to address questions asked by supervisors and fleet managers.

Finally, the PPDB must be explicitly linked to the PADB. If particular deficiencies are revealed in the PPDB they should point to specific aspects of training and curricula in the PADB. This linking can be used at the aggregate level (e.g., to provide additional training of a specific skill to an entire fleet) or it can be used in debriefing an individual pilot or crew. This explicit linkage between the PPDB and PADB is an integral part of AQP.

The AQP Model Database

The cost of developing the PADB and PPDB can be prohibitive for many smaller carriers. To reduce this cost, the FAA is supporting the development of a Model PADB that will be available to all carriers at no cost. At the time of the writing of this document a model PADB is available in Paradox. An ACCESS version of the Model AQP incorporating both a PADB and PPDB will be available in February 1999. A sequel server version, with full functionality will be available in February 2000.

Need for an I/E Database

Because the various measures reflecting I/E proficiency will be unique from those describing pilot/crew proficiency carriers may want to develop a separate database for storing and analyzing I/E performance. We have developed an I/E calibration tool in Microsoft Access that facilitates data collection, analysis, and report generation for LOE sessions. The database that was developed for this software can serve as the beginning of an I/E database. This software will soon be extended to address I/E ratings of First Look and Maneuvers Validations as well. The I/E LOE software is currently available on the FAA Web site and the industry SOP Web site. As soon as the software for First Look and Maneuvers Validations is ready, it will also be available to carriers at no cost at the same internet addresses.

Return to Top

Section III: Training and Assessment of I/Es

In Section I, we stressed that a viable AQP program requires the collection of quality performance data. We need to accurately measure the proficiency of aircrews in order to know whether training is effective and whether certain skills need additional or possibly a different type of training. Ideally, we would be able to measure crew performance with some device that automatically recorded and scored all of the relevant parameters related to the operation of the aircraft. FOQA is an important step in this direction in its ability to measure both performance outcomes and the more technical aspects of flying. However, FOQA will never replace the

expert human observer when it comes to assessing the more behavioral and cognitive aspects of flying. Here, as with certain technical aspects of pilot/crew performance, we must continue to rely on I/E judgments. Therefore, the implementation of AQP is critically dependent on the accuracy of I/E judgments of pilot/crew performance. This requires that every carrier in AQP have a training and calibration program that ensures the proficiency of I/Es. Below we discuss some of the factors that need to be considered in conducting I/E calibration sessions.

Design of Gradesheets

All forms of performance evaluation require some type of a gradesheet. Our focus here is with LOE gradesheets, which assess both technical and CRM types of skills. Later we will briefly discuss the grading of First Look Maneuvers and Line Check Performance. The design and contents of an LOE gradesheet is a reflection of a carrier's philosophy regarding CRM. Given the effort invested in training I/Es to use the gradesheet proficiently, it is not something that a carrier would want to change with any frequency. It is with these thoughts in mind that we encourage carriers to carefully consider the all of the parameters of their LOE gradesheet before committing to a particular design.

Need for both Objective and Subjective Assessment

A recent trend in the design of LOE gradesheets is to require I/Es to grade a flight scenario both in terms of event sets and more specific items (referred to variously as OBs or performance indicators) within event sets. For example, the event set might be an approach under certain abnormal conditions and one of the OBs might be "Captain verbalizes flying/monitoring duties after abnormal correctly identified."

As we commented earlier, the grading of the OBs can provide a more detailed understanding of the factors underlying a weak grade on the overall event set. The OBs also provide pointers or cues as to what the I/E should look for in the grading of a specific event set. This should increase the uniformity in grading of the overall event set. Finally, and perhaps most importantly, the pattern of ratings on the OBs will provide a more specific diagnosis of any problems manifested with a particular event set. With this information we are in a better position of knowing what specific segment of training and curriculum needs to be reviewed.

Some I/Es have argued that the information obtained in OBs is redundant with other types of measures such as written comments or reason codes. Reason codes are a fixed set of reasons that could underlie poor performance on an event set (e.g., poor communication, poor workload management, etc). While we suspect that there is a good deal of validity to this claim, both written comments and reason codes have certain problems associated with them. The most obvious problem with written comments is that they are difficult to quantify and aggregate once they are entered into a database. Reason codes can easily be quantified and may provide an important source of data on LOE performance. However, it is often unclear what specific actions or non-actions within the event set the reason code refers to (e.g., if the reason code "poor communication" is checked for an event set, we do not know if this was true for the entire event set, or if there was a critical event where communication was poor.)

Another common objection to grading OBs is that it may require more work for I/Es, who are already are in a high workload situation running the simulator. While this seems to be a reasonable criticism, it has been our experience that when I/Es become familiar with the event sets and the specific OBs they are grading for each event set, they no longer find the grading of OBs all that demanding.

The Relationship between Event Set and OB Grades

Although a case can be made that the total set of OB grades for an event set should determine the overall event set grade, we believe this would be a mistake. The small numbers of OBs that are listed within an event set are not an exhaustive depiction of the entire event set. Thus, it is always possible that some important unique behavior (either appropriate or inappropriate) occurred that influenced the quality of the overall performance, but was not represented by an OB. This behavior would not be reflected in the OB grading, but it should have an influence on the overall event set grade. For this reason, I/Es should be allowed to grade an event set independently of how they graded the OBs.

A significant discrepancy between an event set grade and the average grade across all of the OBs for that event set is perhaps most likely to occur when a pilot performs well on the OBs, but then makes a critical error that is not covered by any specific OB (i.e., high performance on OBs does not imply high performance on the event set). It seems far less likely that a pilot who scores consistently low on OBs would score high on the event set (i.e., low OB performance usually implies low event set performance). On average we would expect to find a high relationship between event set and OB performance. Thus, the set of OBs for a particular event set should be re-evaluated if it is found that they consistently fail to agree with the event set grade. In particular, the event set may fail include OBs that are important to I/Es' determination of an overall event set grade.

Adopting a Uniform Grading Scale

It is possible to spend an inordinate amount of time deciding on the grading scale to be used for rating performance on different measures. We have seen scales range from 2-point to 5-point ratings. Although there are several factors that could influence the choice of a scale, we are going to limit our discussion to three basic considerations. First, we want our grading scale to capture the full range of variations that I/Es observe in pilot/crew performance. If we simply grade all levels of performance into pass or fail, when in fact the I/Es can reliably discriminate five levels of performance, we are losing potentially valuable information. Therefore, the carrier needs to know how finely the I/Es can reliably discriminate and grade each type of performance (e.g., event sets or OBs). Our analyses of event set ratings data suggest that I/Es can reliably discriminate among three or four levels of acceptable performance. This would dictate the adoption of at least a 4-point scale (i.e., a scale, where there are three levels of acceptability and one level that is not acceptable). We have found that when there are only two levels of acceptable and one non-acceptable level (e.g., acceptable, debrief, and repeat) that virtually all of the performances are graded as acceptable. Our goal in assessment is to combine the finest level of discrimination with the highest possible level of reliability. Within reason, the finer the level of discrimination the better we can determine which aspects of the curriculum are working.

Second, if at all possible, it is preferable that all types of performance be graded on the same grading scale. Having I/Es use a 3-point scale on one task and a 4- or 5-point scale on another task can be confusing, not only to the I/Es, but also to managers when evaluating performance reports. The argument is often made that the grading scale must be adapted to the particular performance that is being graded. In the case of LOE grading, some carriers believe that OBs can only be graded on a 3-point scale (e.g., acceptable; debrief; or repeat), whereas an event set can be graded on a 4-point scale (exceptional; good, pass, unsatisfactory). However, other carriers use a 4- or 5-point scale for both event set and OB grading, often using the same labels regardless of what is being measured (e.g., event sets, OBs, or maneuvers validation). In most cases we have found that it is possible to adopt a single 4- or 5-point grading scale to assess all aspects of aircrew performance. We believe that the advantages of a uniform grading scale far outweigh the potential confusion resulting from using multiple scales.

Third, when rating performance on a scale that goes from unsatisfactory to excellent there is

sometimes a tendency for raters to overuse the middle of the scale (e.g., a rating of 3 on a 5-point scale) when the middle value is considered average or acceptable. A simple examination of the distribution of I/E ratings will reveal whether this is a problem. Should it be found that almost all of the ratings are 3's on a five point scale, it raises the question of whether I/Es are accurately discriminating among levels of performance. Although, a preponderance of middle grades doesn't necessarily indicate a problem (it is possible that this is a valid reflection of performance within a fleet), it would certainly seem to call for a calibration session in which I/Es' ratings are compared with referent based assessment of LOE performance. If there is a consistent tendency for I/Es' to grade a performance as a 3, when the referent grade is a 2 or a 4, it would indicate that I/E's are not using the grading scale appropriately (for more detail on this issue see Johnson & Goldsmith, 1998, pp. 11-13).

One simple means of preventing I/Es from overusing the middle rating is to use an even numbered point scale (e.g., a 4- or 6-point scale) where none of the rating values is labeled average (e.g., on a 4-point scale a 2 may be labeled as slightly below average and a 3 is slightly above average). Of course, it must be recognized that the underlying problem may be far more serious than something that can be addressed by simply changing the number of scale values. We will return to this issue later when we discuss referent based grading.

Grading both Technical and CRM Skills

Whereas the assessment of First Look and Maneuvers Validation is primarily concerned with a pilot's technical skills, LOEs are focused on assessing the integration of technical and CRM skills in the context of a realistic flight scenario. Although specific grading of CRM skills is not required; most carriers that are involved in AQP will want to assess CRM proficiency for purposes of curriculum validation.

How can this type of integrated assessment best be accomplished? It may be argued that event set grading itself accomplishes this by the fact that the overall level of crew performance on an event set necessarily reflects the integration of technical and CRM skills. The assumption being that if the crew was weak on the CRM skills that were of primary importance for a particular event set, this weakness would manifest itself in technical deficiencies and overall weak performance on the event set. While this may be true it must be recognized that the event set grade fails to reveal the specific antecedents that were responsible for the overall event set grade. If the OBs are carefully selected to reflect the specific kinds of CRM skills that are critical to the successful completion of an event set, then they can serve an important diagnostic function. They should reveal the missing antecedents that were largely responsible for any sub-par performance on an event set.

It has been our experience that the content of the OBs for any given event set is usually a mix of technical and CRM oriented items. This allows for the possibility of examining in greater detail the relationship between CRM skills and technical performance (e.g., does failure to properly manage workload increase the likelihood of certain technical deficiencies). Ideally, an examination of the pattern of performance across the entire set of OBs contained within an event set will better reveal the nature of any underlying deficiencies in CRM skills. These kinds of information can greatly improve the effectiveness of LOE debriefings. With this information the I/E is able to point out explicitly how a specific CRM deficiency resulted in a specific technical problem.

I/E Buy-in to the Assessment Process

Because the I/Es play a critical role in the assessment of pilot/crew performance it is important that they have confidence in the methods and tools they use to make their assessments. Toward this end we recommend that they be brought into the development of all assessment

methods. An obvious example of this is in the design of the LOE gradesheet. Simply on the basis of human factors considerations alone it is essential to obtain their input on the design of the gradesheet. Not surprisingly, we have found I/Es to provide thoughtful and valuable information on the design of an LOE gradesheet. Moreover, when conducting a calibration session with a group of I/Es, we have found that they are far more willing to accept feedback on their performance (e.g., how reliable their ratings were) when they had a hand in the design of the gradesheet.

Before leaving this topic, it is important to minimize any potential conflict that may arise between I/Es and supervisory or professional staff in the design of a gradesheet. Clearly, I/Es will evaluate the design of a gradesheet from a somewhat different perspective than a PhD in statistics and test design. I/Es are more likely to be concerned with the ease of entering their ratings while conducting a full flight simulator session, than a supervisor or researcher who may be more interested in obtaining maximally discriminative ratings. We believe that some of this potential conflict can be minimized if the professional staff clearly set forth the minimal properties of the gradesheet before soliciting I/E input. This approach is far preferable to giving I/Es unconstrained freedom in the design conditions and then coming back later to say that most of their recommendations failed to meet some previously unspecified criteria. Based on our experience, the importance of these considerations cannot be understated with regard to achieving acceptance of the AQP performance assessment process.

Appreciation of the Training Implications of Assessment

As we have commented repeatedly, AQP requires that assessment outcomes not only play an evaluative function (is the pilot qualified to fly), but the outcome must also play a training function. The training function is more concerned with an assessment tool's ability to diagnose specific deficits which, in turn, point to particular curriculum and training that are designed to address the deficit. With AQP it is important that I/Es have a clear vision of their role in both the evaluative and training side of assessment.

In our discussions with I/Es, both in the context of designing gradesheets and in conducting calibration sessions, it became apparent that they have a clearer view of the role they play in evaluation than the role they play in training. Their contribution to training is seen primarily in the context of debriefing. When asked to critique an LOE gradesheet they were more inclined to focus on the evaluative function than its potential value as a training device. This is quite understandable, considering the fact that an I/E's evaluation can have a major influence on a pilot's career. As a consequence they often had excellent suggestions on improving its evaluative function. However, their suggestions were usually less concerned with the training implications of the instrument (i.e., will performance on this event set and the associated OBs accurately diagnose the specific skill deficit that underlies marginal performance and point to that part of the training curriculum that would address this weakness). It is reasonable to expect that with the implementation of AQP I/Es will begin to think more critically of an items diagnostic function as it relates to the curriculum and qualification standards as stated in the PADB. It is with this perspective that I/Es will be able to provide more comprehensive input on the design of a carrier's assessment tools.

Return to Top

Section IV: Conducting Group and Individual I/E Calibration Sessions

In this section we discuss those factors that we have found to have a notable affect the quality of a calibration session. Although calibration sessions may be conducted on individuals, as well as groups, our focus will be on group calibration sessions for the simple reason that they are inclusive of most of the issues that arise in conducting individual calibrations. A group calibration session consists of three phases:

1. Data Collection Phase - where the I/Es view and grade a video of a crew performance (e.g., in the case of an LOE it may be a video of a flight scenario that is often segmented into event sets or phases of flight)
2. Data Analysis Phase - where the gradesheets are collected and taken off by a data analyst for statistically analysis and generation of group and individual reports; and
3. Feedback Phase - where the results are presented and discussed with I/Es. In the discussion that follows we will elaborate on several issues related to the above three phases of calibration.

Scripting an LOE video

There are basically two approaches to creating a video for the purposes of calibration. One approach is to video tape an actual LOE and the other is to enact an LOE. There are pros and cons to both approaches. The most obvious problem in using an actual LOE is the difficulty in obtaining permission from the crew and the pilots' union to use the video for training. This is perhaps more likely to be a problem when the video depicts performance that is weak or unsatisfactory. Of course, this problem is completely circumvented by using enacted videos. A potential criticism with enacted videos is that they lack realism. While this is certainly possible, it has been our experience that carriers are quite capable of creating realistic videos. To do so requires a good deal of care in the planning, taping and editing of the video. The bottom line is that the videos must appear realistic to the I/Es. As a matter of course, it is recommended that I/E judgments of video realism be collected as part of the calibration session.

There also seem to be two different philosophies regarding the temporal continuity of the video. One approach is to create a video that follows the natural temporal order of events from preflight planning to the conclusion of the flight at its destination. This type of video may be analyzed into phases of flight, with an event set grade assigned to each phase of flight, but the video is continuous from beginning to end. At the other extreme, the video may begin at any point in the flight and presents some number of event sets (e.g., approach and landing). The successive event sets are usually continuous in the sense they are part of the same flight, but this is not necessarily the case (i.e., the LOE may consist of three discrete and independent event sets). With a discontinuous video it is necessary for the supervisor to provide the I/Es with the relevant background context (e.g., that this is a part of a flight from Albuquerque to New York, and the plane is currently 300 miles from New York, cruising at 31,000 feet, etc.). Also, with this type of video the I/Es are expected to grade all the items related to the current set before moving on to the next event set.

Coordinating the Gradesheet with LOE Video

The coordination of the gradesheet with the video is a relatively trivial matter when a carrier only grades event sets. However, if the gradesheet calls for grading OBs it is important that the specific behaviors that are referred to in the gradesheet are clearly presented in the video. If the video is scripted to omit a specific behavior it should be clear by the context of the video, where the behavior should have occurred. Once the video and the gradesheet are developed it is necessary to have a group of I/E supervisors observe the video and determine that there is no ambiguity as to whether an OB did or did not occur.

Assigning Referent Grades based on Qualification Standards

We recommend that LOE grading be based on what we refer to as a referent-based approach (Johnson & Goldsmith, 1998, pp. 7 - 11). The philosophy behind a referent-based approach is that a pilot's performance is evaluated relative to the air carrier's qualification standards and the appropriate application of the air carrier's grading scale. The assumption underlying this approach is that there is a specific grade that should be assigned to a specific level of performance. Applying this approach to I/E calibration training requires that a referent grade be assigned to each item on the gradesheet. A referent grade can be assigned by having a group of four to six supervisory I/Es apply the qualification standards and the grading scale to arrive at a rating for each item. Any differences in grading among the supervisors should be resolved by having the supervisors review the appropriate segment of the video and discuss the basis for the differences in grading. If the differences are not easily resolved it may suggest that (a) the relevant performance is not clearly presented in the video, (b) the item is not clearly stated on the gradesheet, or (c) the link between the item and the appropriate qualification standard is not clearly defined. If these problems cannot be resolved then the item should be removed from the gradesheet.

Preparing I/Es for grading a LOE video

Prior to conducting an I/E calibration session we recommend that the participating I/Es be given an opportunity to review the relevant qualification standards. This may be accomplished by simply mailing a copy of the gradesheet that is to be used in the upcoming calibration session to the I/Es one-week before the session.. The I/Es can be alerted to the fact that they should be familiar with the qualification standards pertaining to the enclosed gradesheet. One might object that this advanced information unfairly prepares the I/Es for the calibration session and as a result their performance is artificially elevated. However, we would contend that the goal is to obtain an estimate of I/E performance as it normally occurs in the simulator. Here, I/Es should always be highly familiar with the upcoming event sets and the corresponding gradesheet.

Collecting I/E Ratings

Upon arriving for the calibration session the I/Es are first given a copy of the gradesheet and then given a complete overview of the three phases and specifically what they will be expected to do. To maximize buy-in we recommend that a supervisor I/E who is well known and highly respected by the I/Es conduct the overview. Given the level of turnover of I/Es within the fleets one can expect there to be a wide range of skills and experiences represented. With this in mind it may be prudent to err in the direction of assuming too little in the way of I/E background knowledge. In this regard, we recommend that the overview include a review of the grading scale and the event sets that they will be viewing. As part of the overview it is important to emphasize that the I/Es are expected to make their ratings independently. When conducting a calibration session with a large group of I/Es it is very easy to lose control, with numerous clusters of I/Es carrying on discussions while the video is being shown. This, of course, will invalidate the results from the session. At the very least, the individual ratings can no longer be viewed as independent on one another.

At the conclusion of the overview the I/Es provide whatever demographic information is required on the gradesheet. Typically this includes their name, an ID number, type of fleet (if all I/Es are not from the same fleet), and whether they are a captain or first officer. On occasion, for research purposes, it may be desirable to collect additional demographic information, such as, year's experience as a commercial pilot, experience as an I/E, type of qualification as an I/E, etc. In these instances when more detailed demographic information is desired it may preferable to simply attach a demographics form that is appended to the gradesheet.

When the video is comprised of more or less discrete event sets, the I/Es maybe told that they

have the option of either entering their ratings as the video is playing or that they may take notes and enter their ratings after the event set has been presented (assuming there is no carrier or fleet policy on when gradesheets are completed in the simulator). After the first event set is presented there is a pause until all I/Es have entered their ratings and then the next event set is introduced and presented. This process continues until all of the event sets have been presented. If the video is more continuous, proceeding through the various phases of flight in the temporal order of a normal flight, there may be no need to pause the video between phases of flight. In this case the I/Es should be told to enter their ratings as though they were conducting an LOE. Keep in mind that in the calibration session we are trying to estimate how I/Es perform on-the-job.

When I/Es have completed making their ratings, the gradesheets are collected for scoring and statistical analysis. Depending upon the size of the group this entire process usually can be completed within two hours.

Measures of I/E Performance

There are two dimensions of I/E grading performance that need to be assessed. One reflects the degree to which an I/E's ratings are sensitive to changes in performance levels (i.e., as performance improves, the ratings go up, and as performance diminishes, the ratings go down). The second factor, what we refer to as scale accuracy, reflects how appropriately the levels of the grading scale are used. These two aspects of performance, sensitivity and scale accuracy, are partially independent in the sense that it is possible for an I/E to be highly sensitive, but poor in scale accuracy. For this reason both of these measures of I/E performance are required. In our paper on quality data (Johnson & Goldsmith, 1998, pp. 11-13) we discuss in more detail two methods of quantifying sensitivity (inter-rater reliability (IRR) and referent-rater reliability (RRR)) and one method of quantifying scale accuracy (mean absolute deviation (MAD)).

IRR and RRR are both based on a correlation statistic. In the case of IRR an I/E's ratings are correlated with each of the other I/Es' ratings and these correlations are then averaged. RRR is computed by correlating each I/E's ratings with the referent grade. While IRR and RRR are usually quite highly correlated with one another, indicating that the two measures of reliability are closely related, they do provide somewhat different information. IRR directly tells an I/E how closely his grading corresponds to the other I/Es. This is important information for an I/E. RRR, on the other hand, informs supervisors how closely I/Es' ratings correspond to qualification standards. This has important implications for curriculum design and training (e.g., it describes where most I/Es need additional training on qualification standards). Because of these differences, most carriers will likely decide to use both the RRR and IRR measures of sensitivity.

MAD is an extremely simple and direct measure of scale accuracy. It is computed by taking the absolute difference between the I/E's rating on an item and the referent value for the same item. This is done for each item and then an average is computed across items. There are some finer nuances to MAD that are discussed in the Johnson & Goldsmith, 1998, pp. 11-13 document; however for the present purposes it is only important to recognize that MAD is a measure of I/Es' appropriate use of the grading scale. MAD will tell you if an I/E is consistently grading too high or too low.

Computing RRR, IRI, and MAD

The paper we referred to earlier (Johnson & Goldsmith, 1998, pp. 7 - 11) provides a detailed discussion of how RRR, IRR, and MAD are computed. The actual computation of these statistics can be done with the calibration software that we developed. The calibration software

will compute these statistics automatically once the data for a group of I/Es has been entered into a data file.

Generation of Statistical Reports

The calibration software, referred to above, generates both group and individual I/E reports. The group report presents the RRR and IRR correlations computed across OBs and the average MAD for each of the event sets. In addition, it presents the MAD for each OB, rank-ordering the items in terms of the group average MAD. This allows for easy identification of those items creating the greatest degree of disagreement. The individual report presents a summary report for each I/E, containing his/her performance on the event sets and the OBs. For event sets, it presents the I/E's ratings compared to the referent rating for each event set. For OBs it presents the I/E's RRR, IRR and MAD. Finally, it presents his/her rating on each OB compared with the referent rating and the group average rating.

Calibration Feedback to I/E

It is in the feedback phase of calibration where the most important I/E training is likely to occur. At this time the statistical reports that were described above are distributed to I/Es and interpreted by the supervisor conducting the calibration session.

To derive the maximal training benefits from the feedback it is important that the I/Es see the calibration session more as a training experience than as a personal evaluation of their performance. In this regard, care should be taken that the feedback is not presented in a manner that makes it appear as a critique. There are various ways in which the appropriate tone for the feedback session can be set. One suggestion is to begin with a discussion of the most positive aspects of I/E performance. For example, attention could first be focused on those OBs with the smallest MADs (i.e., the highest level of agreement with the referent). Typically there are several items where there is perfect agreement across all of the I/Es. This feedback also demonstrates that it is possible to attain very high levels of agreement in their ratings. When turning to the items with the lowest level of agreement they can be discussed from the perspective of what can be done to improve these items. With these items it may be useful to replay that segment of the video pertaining to specific low agreement items. Typically this will evoke a discussion in which several of the I/Es will comment why they scored that particular item higher or lower than the referent. This discussion can be one of the most important aspects of the feedback phase. All I/Es, even those who choose not to participate in the discussion, will gain an appreciation of the different types of mistakes that are possible. The overall impact of the feedback is dependent on I/E buy-in. If I/Es do not have confidence in the process, they are likely to ignore or rationalize the information contained in the feedback (e.g., it wasn't me, it was the gradesheet, or the video, etc.).

Before concluding our discussion of the feedback phase it is important to note that feedback is a two-way process. In addition to supervisors presenting feedback to the I/Es, it is of equal importance that the I/Es provide feedback to the supervisors and designers of the video and the gradesheet. At a global level it is important to obtain I/Es' overall perception regarding the quality of the calibration session (e.g., what did they benefit most from, what was the least beneficial, how could it be improved, etc.). At a more detailed level we should solicit feedback regarding specific items on the gradesheet and specific events contained in the video. It is only through this type of process that the quality of the videos and the gradesheet can be continually improved.

Linking Feedback to Qualification Standards

During the feedback phase, particularly during the analysis of performance on individual OBs,

it is important to have ready access to the qualification standards that pertain to each item. This provides the basis for what is the appropriate referent grade for each item on the gradesheet. Our calibration software is capable of storing the relevant qualification standards for each OB. The supervisor can simply 'point and click' an OB on the computer screen and the qualification standard for that item will appear. This can be an important tool in teaching I/Es how qualification standards are translated into a referent grade for an item. It explicitly illustrates how reliable and valid grading requires knowledge of the qualification standards.

Conducting Individualized LOE Calibration Sessions

The calibration software that we have developed allows an air carrier to conduct individualized I/E calibration sessions. Here an I/E can login at a workstation, view an LOE video on the monitor, enter his/her ratings on the computer-displayed gradesheet, and receive complete feedback on his/her performance.

Our goal was to replicate as closely as possible what occurs in the group calibration session. While the individualized calibration session cannot replicate the verbal interaction that normally takes place during the feedback phase of a group session, it does have several unique advantages: 1) it avoids the logistics of arranging for a group session, an I/E can complete a session whenever he/she has a free hour; 2) the selection of the event sets can be customized to the I/E's personal training history; 3) some or all individualized sessions can be structured exclusively as training experience, essentially allowing for self training for upcoming group calibration sessions; and 4) the feedback would be completely individualized, allowing the I/E to replay parts of the video where he/she deviated from the referent and to review the relevant qualification standards that pertain to a specific OB or an entire event set.

This software is currently available on request by contacting either author. To implement the software, each carrier would have to develop the necessary videos, and the corresponding gradesheets with referent ratings. Of course, to provide feedback the specific qualification standards pertaining to each item on the gradesheets would also have to be entered into the appropriate file.

Extending Calibration to First Look/Maneuvers Validation

Shortly, we will be extending the LOE calibration software to include First Look and Maneuvers Evaluation. We expect to have this software ready for distribution to carriers by the end of August 1999. The major effort here will involve the creation of a set of videos showing the various critical maneuvers that are assessed on First Look evaluations. Thus, if a fleet uses ten different critical maneuvers (e.g., rejected take off, V1 cut, etc.) they may have videos showing at least two levels of performance on each of these maneuvers. In most cases, each critical maneuver would last no longer than a couple of minutes; therefore the entire video would likely not exceed 20 minutes in duration. The most obvious difference between an LOE and a Maneuvers Validation video is that that the information presented in a critical maneuver would focus far more on instrument displays, whereas most of the information in an LOE video is more contained in the verbal interactions and physical actions taken by the crew.

Once the videos were created the software would allow carriers to calibrate I/E evaluations of technical maneuvers in the same manner that LOE evaluations are calibrated. The same three statistics (RRR, IRR, and MAD) used on LOE ratings could be computed on First Look ratings. For example, in a group calibration session I/Es could be shown 20 critical maneuvers, grading each on a 4-point scale. The I/E supervisors would establish a referent evaluation for each of the 20 maneuvers and this would be used to compute RRR and MAD. IRR would be computed as before, by simply correlating each I/E's ratings with the ratings of each of the other I/Es in the group and then calculating the mean of these correlations. The calibration feedback would

proceed in basically the same manner as described for an LOE session.

Extending Calibration to Line Checks

Although we have not begun to develop the software to conduct Line Check calibration sessions, there is no reason why the same approach could not be used in this context. Producing a video of a simulated line flight should be quite similar to an LOE. The most obvious differences between an LOE and Line Evaluation are:

1. a line flight is not typically segmented into event sets (however, both can be analyzed in terms of phases of flight)
2. a line flight does not contain specific triggering events creating abnormal flying conditions
3. a line flight is usually of longer duration
4. a line flight is not necessarily evaluated in terms of a grade sheet with specific items (e.g., OBs).

The primary concern in conducting a line evaluation calibration would be to make it as similar as possible to the way line checks are normally conducted using the same grading instruments, and making the videos as representative as possible to what occurs in actual line flights. The only engineered deviation from normal line flights would involve scripting levels of crew performance to ensure some degree of variation in performance levels and possibly editing out segments of the cruise phase of the flight to shorten the duration of the video.

Return to Top

Section V: Evaluation and Training of Pilots and Crews within AQP

What has changed with AQP?

As we emphasized at the beginning of this document, the ultimate concern of AQP is the training of pilots/crews to proficiency, where proficiency is defined in terms of qualification standards. All of the concern with I/Es' assessment tools, statistical measures, etc., should be viewed as only as a means to ensuring pilot/crew proficiency. Therefore, what has changed is that a pilot's competency is no longer defined simply in terms of the hours, days, or years of training he/she has received, but in terms of his/her current proficiency. When properly implemented this should reduce or eliminate training of skills that are already highly proficient, while at the same time increasing the training of specific skills that are in need of training.

Over time, with the accumulation of reliable data in the PPDB and the appropriate statistical analyses of these data, it will be possible for carriers and the FAA to empirically determine how often and to what degree, specific skills need to be trained. This introduces the possibility of a win-win situation, where the carrier is able to reduce expenditures for unnecessary training and the pilots are uniformly more proficient.

Defining and Measuring Proficiency

Again, with the implementation of AQP and the accumulation of data in the PPDB, the precision and quantification of different skill proficiencies will be continuously sharpened. For example, with the appropriate statistical analyses of the PPDB and the PADB it will become increasingly clear what training will best address specific skill deficits. In short, the assessment

methods will continue to improve, which in turn will improve the efficiency of training content and methods. This is good news for the pilot, the management, the FAA, and most importantly the flying public.

Relevant Documents

Data Management Guide (5/12/98) Produced by the Air Transport Association's Data Management Focus Group

The Importance of Quality Data in Evaluating Aircrew Performance, FAA , Peder J. Johnson and Timothy E. Goldsmith, Dept. of Psychology, University of New Mexico, Albuquerque, NM, 87131

Questions Your Database Should Address, Peder J. Johnson & Timothy E. Goldsmith, Department. of Psychology, University of New Mexico, Albuquerque, NM, 87131.

Contacting Authors

Dr. Peder J. Johnson
Department of Psychology
University of New Mexico
Albuquerque, NM 87131
(505) 277-4339
email: pjohnson@unm.edu

Dr. Timothy E. Goldsmith
Department of Psychology
University of New Mexico
Albuquerque, NM 87131
(505) 277-7505
email: gold@unm.edu