Analysis of Training of Cognitive Skills in a Line-Oriented Flight Training Environment

Timothy E. Goldsmith and Peder J. Johnson
Department of Psychology
University of New Mexico

The following document constitutes the final report for FAA Grant 94-G-013, Analysis of Training of Cognitive Skills in a Line-Oriented Flight Training Environment. The focus of this research was on improving the assessment and training of skills and knowledge of airline pilots. A major theme of the work attempted to validate the assessment of Crew Resource Management (CRM) skills, an area that had previously been identified as relevant to airline accidents and incidents. Our results showed that current methods of assessing CRM skills using observed behaviors might not adequately discriminate among performance in these skill categories. The problem may lie with the validity of the skill categories themselves, the adequacy of observed behaviors to reflect CRM skills, or the ability of evaluators to discriminate among different levels of CRM performance. Second, our research evaluated the psychometric properties of aircrew assessments, and the results indicated several areas where changes to the assessment methods could potentially benefit the quality of aircrew assessments. These areas included design of gradesheets, wording of performance indicators, grade scale labels and number of levels, and training of the evaluators themselves. To support this last recommendation, we developed of a software package to aid in training and calibrating instructors and evaluators. The software has been made available to all major US carriers.

The Importance of Quality Data in Evaluating Aircrew Performance

INTRODUCTION

In this paper, we discuss the application of basic psychometric principles to the problem of assessing aircrew performance. In particular, we are concerned with evaluating aircrews under the *Advanced Qualification Program* (AQP) in high fidelity, full-flight simulators. A major goal of AQP is to provide the carrier with a *quality assurance* program which ensures that aircrew members have the highest possible level of proficiency on all technical and management skills relevant to the safe and efficient operation of the aircraft. The implementation of a quality assurance program requires a database system that begins with an explicit set of qualification standards that are based on job task listings. These qualification standards drive the content of the curriculum, which in turn drive an assessment process that explicitly evaluates pilots on these qualification standards. The data from the assessment process provides feedback regarding the content and delivery of the curriculum. This feedback in turn allows for continuous improvements in curriculum design, as well as better directing the allocation of training efforts to those knowledges and skills that are weakest. When functioning properly this system will ensure that all aircrew members attain and maintain a pre-specified standard of proficiency. Thus, it can be seen that quality assurance requires *quality assessment.*

A quality assurance program can only be as good as its weakest link. The qualification standards must be based on a careful analysis of job task listings. The curriculum and instruction must be designed to train to the qualification standards. And finally, the focus of this chapter, the assessment tools must provide a *reliable* and *valid* evaluation of performance. It is of the utmost importance to realize that under AQP we are not simply assessing individuals, we are assessing the viability of the curriculum, the instructors, and the evaluators. From this perspective, the primary function of assessment is to improve training and thereby provide highly qualified aircrews.

Overview

The primary goal of this chapter is to describe a set of methods and procedures that will enhance the quality of the data used to assess aircrew performance. The two fundamental properties of quality data are reliability and validity. This section begins with a formal discussion of these two ideas, including a description of the statistics used to estimate reliability. After

giving a formal treatment of reliability and validity we next discuss these concepts in the context of aircrew performance assessment. Here our discussion will be concerned with the three primary factors that influence the overall quality of the data. The first is the observer or evaluator who must make the judgments or ratings of the observed performance. The second is the measuring instrument (e.g., a Line-Oriented Evaluation [LOE] grade sheet) that is used to collect data. The third factor is the host of parameters that comprise the assessment situation (e.g., a calibration session). As a brief aside it is important to understand that the assessment situation is often not the same situation under which assessments are normally conducted. For example, in a calibration session the evaluators will observe and judge a video of a crew flying an LOE as opposed to judging an LOE simulated flight. This is necessary because in order to estimate reliability every evaluator must observe the identical crew performance. The video is necessary because it would pose some obvious logistical problems to arrange for 20 or more evaluators to observe an actual LOE in the simulator. Returning to the central point of this discussion, when we refer to the parameters of the assessment situation, it must be understood that they are not always the same as the conditions under which these types of observations are normally made.

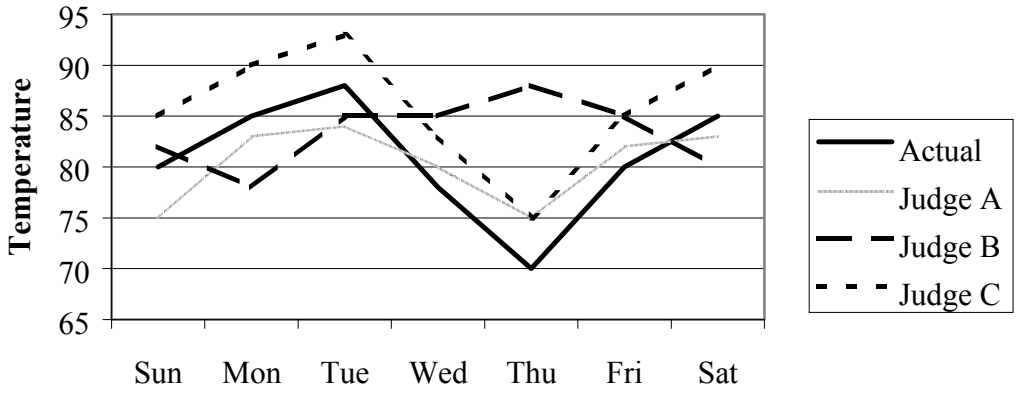

Figure 1. Graphs depicting different Judge's deviation from actual temperature

RELIABILITY

Reliability is a concern whenever we are engaged in observation or measurement. It is concerned with the *consistency* of our measurements (Anastasia, 1958). Thus, if we repeatedly weigh the same brick and we observe little or no variation in the outcomes, we would conclude that our

observation or measurements are reliable. In this chapter, we want to extend the definition of reliability to include the properties of *sensitivity* and *accuracy*. Sensitivity refers specifically to the degree to which observations track or covary with changes in the object that is being measured. The concept of sensitivity is depicted in Figure 1. Judges' estimates of the high temperature at Atlanta airport over a seven-day period are compared with the actual temperature as recorded by the US Weather Bureau. We see judge A's estimates covary very closely with the true temperature, whereas judge B deviates almost randomly from the true temperature. We would conclude that judge A is more sensitive to temperature variations than judge B. In this sense it can be said that judge A is more reliable than judge B.

Accuracy refers to how closely our measurements correspond to the absolute magnitude of what is being measured. This idea is illustrated in Figure 1 by judge C, who is extremely sensitive to actual variations in temperature (i.e., his estimates covary precisely with the true temperature), but who consistently overestimates the temperature by 5 degrees. In this regard, we would conclude that judge C is not highly reliable. In sum, reliable measurements must be both sensitive and accurate.

Quantifying Sensitivity: Correlational Measures

Although there are a variety of statistics for quantifying the degree of sensitivity in a measurement, the Pearson product-moment correlation (indicated by *r)* is by far the most commonly used. Its popularity with researchers and statisticians is due to several desirable properties. First, the Pearson correlation varies on a continuous scale between the values of -1.0 and +1.0. The direction of the correlation or relationship is indicated by the plus or minus sign to the left of the numerical value. A positive relation occurs when the two variables covary in the same direction (e.g., as individuals' height increases they are also likely to weigh more). A negative correlation depicts an inverse relationship (e.g., as altitude increases the content of oxygen in the atmosphere decreases). These two types of relationships are depicted in Figure 2.

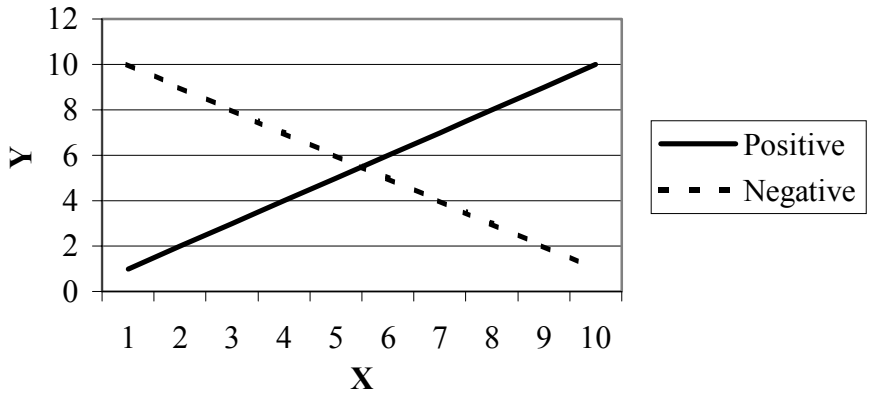

Figure 2. Illustration of positive and negative correlations

Second, as the magnitude of the correlation moves from 0.0 to either +1.0 or -1.0, the strength of the relationship increases. The idea of strength of relationship can be illustrated with a *scatter plot.* Figure 3 plots the relationship between height and weight,

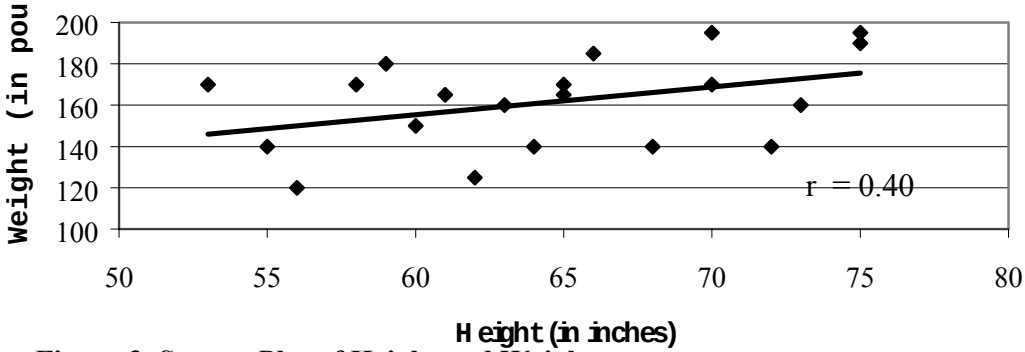

Figure 3. Scatter Plot of Height and Weight

where each point represents a single individual's location on the two axes. As can be seen from this scatter plot, for every height there is a range of weights. The variation from a perfect relationship can be seen as the deviations from the straight line drawn through the scattering of data points. The line represents the best linear (i.e., straight line) fit to these data. The relationship between height and weight is obviously imperfect.

The Pearson correlation statistic reflects the amount of scatter or variance we see in Figure 3. As the variance decreases and scores move closer to the best-fitting straight line, the magnitude of the correlation increases. This idea can be seen in Figure 4 which shows the relationship between the distance between two points and the time required for a given type of aircraft to fly the distance. Here we see far less deviations of the observations (points) from the best fitting line. For a given distance there is relatively little variation in times that are observed. The Pearson correlations for the data presented in Figures 3 and 4 are 0.40 and 0.80, respectively. Notice that the slopes of the lines in these two figures are very similar. The difference is in the amount of variability or scatter of the points above and below the lines.

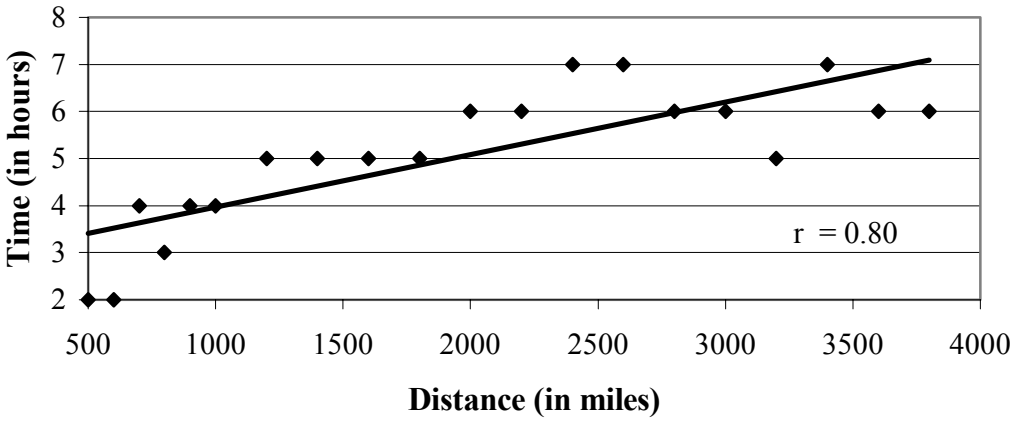

Figure 4. Scatter plot showing covariation between time and distance

Another attractive feature of a Pearson correlation is that the square of the correlation tells us how much of the variation in one variable is accounted for by knowing the other variable. For example, the correlation of 0.40 between height and weight indicates that 16% (0.40^2) of the variance in weight is accounted for by knowing a person's height (or vice versa). Or said differently, the variation in the weight among a group of individuals is reduced by 16% if we were to control for differences in height (e.g., if every member of the group was 72 inches tall). In Figure 4, where the correlation is 0.80, 64% of the variance in time is accounted for by distance. Thus, knowing distance traveled severely constrains the variation in travel time.

Figure 5 illustrates how a scatter plot can be used to depict sensitivity between two judges. Assume that two evaluators are shown a video of a crew flying an LOE and each evaluator independently rates the Captain on the same 20 performance indicators using a 4-point scale. Along the horizontal axis we have the ratings of Evaluator 1 and along the vertical axis the ratings of Evaluator 2. The 20 points on the scatter plot (not all points are visible because of redundancy) correspond to the 20 performance indicators. For example, it can be seen that Evaluator 1 rated some performance indicators a "2", whereas Evaluator 2 rated the same performance indicators "2"s and "3"s. In this way we can see how the scatter plot depicts the

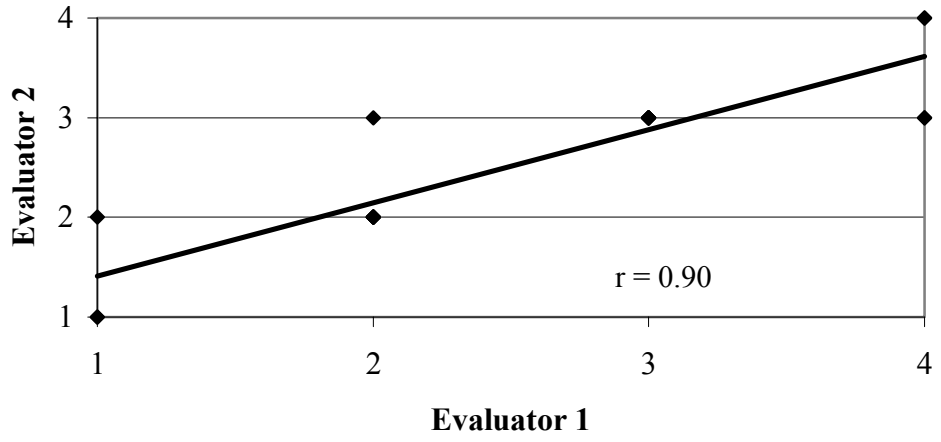

Figure 5. Two Evlauators' ratings of 20 performance indicators on a 4-point scale.

degree of agreement between the two evaluators. Once again, if there were perfect agreement between the two raters, all of the points would fall on the best-fitting straight line across the graph. The Pearson correlation for these data is 0.90.

In a limited sense, the Pearson correlation characterizes the information in the scatter plot with a single statistic. As noted earlier, there are several different measures of correlation and none share all the properties of the Pearson correlation statistic. Therefore, it is important when reading reports containing correlational statistics to know whether it is a Pearson statistic or not.

Two Correlational Measures of Sensitivity

In this section we discuss two methods for assessing the reliability of observations, rater-referent reliability (RRR) and inter-rater reliability (IRR). Although both methods can be said to measure reliability, we believe RRR is the better measure of sensitivity. Also, the reader should be forewarned that these labels (RRR and IRR) are somewhat misleading in that they suggest they are measures of *rater* reliability, when in fact they also reflect the influence of the measuring instrument and various other factors that influence the sensitivity of the observations. These factors are discussed in some detail later in the chapter.

Rater-Referent Reliability (RRR)

RRR is a correlation reflecting how closely an evaluator's ratings agree with some standard or referent. This method of assessing sensitivity can be used when there is an external, objective basis for defining a referent score. A simple illustration is a situation where we correlate an individual's subjective estimates of the weights of different objects with their actual weights. To the extent that the subjective estimates track or covary with the actual weights, the estimates are sensitive and the individual's RRR will be high.

RRR can be used to assess evaluators' sensitivity in assessing aircrew performance as long as we have an objective basis for grading performance. This is the situation at several carriers, where there are explicit qualification standards for grading LOE, First Look, and Maneuvers Validation performance. These performance standards are set forth in the fleet qualification standards and these standards serve as the basis for curriculum and training. In addition, there are clearly established grading criteria that map degrees of deviation from the performance standards onto the grading scale (e.g., was this performance a 4, 3, 2, or 1 on a 4-

point grading scale?). With this type of information it is possible for evaluator trainers to script videos that capture specific deviations from the qualification standards. These videos can then be validated by having a group of evaluator supervisors view and grade the aircrew performance on the video. These expert ratings then become the referent values for computing RRR.

We will make this procedure more concrete by illustrating how RRR can be used to assess the sensitivity of evaluators' ratings of LOE performance. Assume we create a video of a crew flying an LOE and develop a grade sheet containing ten performance indicators (sometimes referred to as observed behaviors) corresponding to specific behaviors that should have been executed according to a task analysis of the phases of flight and the events occurring. Assume further that the video shows the crew deviating from standard operating procedures in a manner that relates to specific performance indicators on the grade sheet. A group of I/E supervisors then grade all of the items on the grade sheet. Any discrepancies among the supervisors are resolved to arrive at a referent value for each of the performance indicators.

At this point we would present the video to a group of evaluators who rate the same performance indicators. As an example of the resulting data, Table 1 shows the ratings for five evaluators along with the referent scores for ten performance indicators. The ratings are given on a 4-point scale.

Table 1. Ratings of 5 Evaluators scoring 10 Performance Indicators

		Evaluators					Referent Score
		1	2	3	4	5	
Performance Indicators	1	2	2	3	3	2	2
	2	1	3	3	3	3	3
	3	4	3	3	3	1	3
	4	4	3	4	3	3	3
	5	3	3	3	3	3	3
	6	2	2	2	3	2	2
	7	2	3	4	3	3	3
	8	1	2	2	1	1	2
	9	3	3	3	3	3	3
	10	4	3	3	4	3	3

It is important to recognize that RRR is a measure of sensitivity because it reflects the degree to which the evaluators' ratings covary with the true performance as defined by the referent rating. As will become clear shortly this is not necessarily true of the IRR measure.

Inter-Rater Reliability (IRR)

IRR is a correlation reflecting the degree to which a group of raters agree with one another. It is the most commonly used method of measuring rater reliability and does not require a referent value. We will illustrate how IRR is computed using the data from Table 1. Each of the five evaluator's ratings is correlated with the ratings of each of the

Table 2. Inter-correlation matrix of five evaluators

		Evaluators					
		1	**2**	**3**	**4**	**5**	**RRR**
Evaluators	**1**	1.00	0.55	0.43	0.59	0.18	**0.55**
	2	0.55	1.00	0.69	0.53	0.60	**1.00**
	3	0.43	0.69	1.00	0.45	0.59	**0.69**
	4	0.59	0.53	0.45	1.00	0.61	**0.53**
	5	0.18	0.60	0.59	0.61	1.00	**0.60**
	IRR	**0.44**	**0.59**	**0.54**	**0.55**	**0.50**	

remaining other four evaluators (e.g., Evaluator 1 with 2, 1 with 3, etc.) resulting in the matrix of correlations shown in Table 2. The bottom row, labeled "IRR" shows the average of these four correlations for each evaluator. The average of these five individual IRRs gives an overall IRR for the group of 0.52. The right-most column of Table 2, labeled "RRR", shows the correlation of each evaluator's ratings with the referent. The overall RRR is computed by simply averaging these five correlations, which in this case is 0.67.

Comparison of RRR and IRR

RRR and IRR are similar in that both are correlational measures reflecting the degree to which measurements covary. However, IRR reflects the covariance between evaluators, whereas RRR reflects the covariance between evaluators and the true score (i.e., the referent). As a result, RRR is necessarily a measure of sensitivity whereas IRR does not necessarily reflect evaluators' sensitivity. One can easily imagine a situation where a group of evaluators is in high agreement with one another (high IRR), but their ratings do not covary with actual changes in the object or event that is being judged (low RRR). This could occur if judges are uniformly basing

their ratings on some irrelevant property of the object being judged. A simple example of this would be young children judging the weight of objects on the basis of volume, rather than mass. Thus the childrens' rating might show high IRR, but quite low RRR.

In most real-world situations, including the evaluation of aircrew performance, we would expect RRR and IRR to be highly correlated with one another. However, while a high RRR implies a high IRR (i.e., if all of the evaluators' ratings are in close agreement with the referent they must also agree with one another), a high IRR does not imply a high RRR. Consider a situation where performance is again being judged on a 4-point scale (4 = outstanding and 1 = unacceptable), but evaluators only use the intermediate values of the scale (3 = acceptable and 2 = minimally acceptable). This could result in high IRR, but the evaluators would be insensitive to the full range of performances being observed, resulting in relatively low RRR.

An additional advantage of RRR over IRR is that it defines a clear objective for training that is based on the qualification standards. With appropriate training on qualification standards and applying grading scale criteria, evaluator's judgments should begin to show high agreement with referent values. Accomplishing this would seem to be an important objective for an airline.

For these reasons we shall consider RRR as the primary measure of sensitivity. IRR can be used as a means of diagnosing RRR values that are lower than expected. For example, if it were found that IRR was higher than RRR and that most of the evaluators disagreed with the referent on a particular performance item or subset of the items, then we would certainly want to resolve the disagreement.

In concluding our discussion of RRR there are three additional points that need to be made. First, although qualification standards contribute significantly to objectifying grading LOE and maneuvers-validation performance, there will always remain a subjective component to the grading process. Even among the most experienced evaluators we may find some degree of disagreement in the assignment of grades (e.g., on a 4-point scale some disagreements between ratings of 2 and 3 are to be expected). However, with training on identifying qualification standards and applying grading scale criteria, deviations of 2 points or greater on a 4-point scale should be virtually eliminated. Later we discuss how calibration sessions can be used to fine-tune an evaluator's application of his knowledge of qualification standards to the grading process.

Second, one reservation regarding RRR is the possibility that a group of evaluators would deviate from the referent for valid reasons. This situation might occur, if for no other reason, because of a clerical error in defining the referent. Computing only RRR would fail to reveal the error. Therefore, we recommend always checking the deviation between the referent ratings and the group's averaged ratings. Significant deviations in ratings of either performance indicators or event sets would signify potential problems to be further investigated.

Finally, a discussion of RRR could easily have occurred in the context of the validity section later in this chapter. Validity concerns the question of whether a measuring instrument truly measures what it is intended to measure and sensitivity obviously relates to this issue. However, in terms of the application of these measures to real situations, we believe that RRR is more closely related to reliability than it is to validity.

Quantifying Accuracy: Mean Absolute Difference

Mean absolute deviation (MAD) is an extremely simple and direct method for estimating the accuracy of observations. It is computed by simply averaging the absolute deviations between the observer's rating and the referent rating as shown in Table 3 for six evaluators' ratings on three LOE event sets. Here it can be seen that a separate MAD was computed across the six evaluators for each of the three event sets. The value of MAD may range from a minimum of 0.0 (all evaluators gave the same rating) to a maximum value that is equal to the difference between the highest and lowest scale value (e.g., on a 4-point scale the maximum MAD would be (4 - 1) = 3). This property of MAD makes it difficult to compare MAD values across different scales of measurement (e.g., 3- versus 4-point ratings). However, the problem is easily rectified by standardizing MAD in term of the number of values on the measurement scale (i.e., MAD divided by the maximum deviation) and then subtracting this value from 1.0. This standardized MAD, referred to as SMAD (See Table 3), ranges between 0.0 and 1.0, with 1.0 indicating perfect agreement. Thus, SMAD allows meaningful comparisons across different scales of measurement and is scaled similar to a correlational measure . For the purposes of the present chapter we will simply use MAD to refer to the generic measure.

Comparing MAD with Sensitivity Measures (RRR and IRR).

In one sense it can be said that MAD is a more fundamental measure than RRR or IRR because if we observe a small MAD we not only know that we have accuracy, but we also have sensitivity. Quite simply, when MAD is equal to 0.00, then both RRR and IRR would necessarily be equal to +1.0. As MAD becomes larger we might generally expect RRR and IRR to approach 0.0, but this is not necessarily the case. As was illustrated in Figure 1, it is possible for correlational measures such as RRR or IRR to be +1.0 when MAD is arbitrarily large. For this reason it is necessary to compute both MAD and RRR to assess both the accuracy and sensitivity of our observations.

Table 3: Calculation of MAD for six Evaluators and three Event Sets

		Evaluators								
		1	2	3	4	5	6	Referent	MAD	SMAD
Event Sets	1	3(0)	3(0)	3(0)	4(1)	3(0)	2(1)	3	.33	.11
	2	3(1)	4(0)	3(1)	3(1)	4(0)	2(2)	4	.83	.28
	3	1(1)	4(2)	4(2)	1(1)	3(1)	4(2)	2	1.50	.50

Note. Under evaluators the digit to the left is the rating given and the digit in parentheses is the absolute difference between the evaluator's and referent rating. MAD = sum of differences divided by number of evaluators (6). SMAD = the MAD value divided by 3.0 (the maximum deviation on a 4-point scale) and subtracted from1.0.

Another potential difference between MAD and correlational measures, such as RRR and IRR, is illustrated in Table 3 where a separate MAD is computed for each of the three events sets. The reliability of the MAD statistic as it is computed here depends on the number of evaluators. This may be contrasted with the RRR and IRR correlational statisitics as they were computed from Table 2. Notice that in Table 2 reliability is computed between a pair of raters (or a rater-referent pair) across the items in the test. The reliability of this statistic is therefore dependent on the number of items rather than the number of raters. This difference between MAD and RRR or IRR may be important when we are attempting to estimate the reliability of instruments containing only a small number of items. For example, if we were assessing an LOE that only contained a few event sets it would not be highly informative to compute RRR or IRR across three or four event sets. However, if we had ratings from a large number of evaluators on

each of these event sets we could compute a reliable MAD value for each of the event sets. It must be recognized that when MAD is used in this manner, it is estimating the accuracy of judgments made on a single item. The inference we are making is to the population of evaluators. This stands in contrast to RRR or IRR which is sampling items and making an inference to some population of items. The point that we are making here is that they are different measures and as a consequence MAD may be used in situations that do not easily lend themselves to a correlational analysis.

Congruency

Before concluding our discussion of accuracy there is one additional measure that can often enhance our understanding of rating data. Congruency is a measure of the degree to which individual raters are distributing their ratings in a manner that is congruent with the referent. Figure 6 shows the frequency with which an individual evaluator used each of the scale values on a 4-point rating scale, compared to the referent. Here we can see that this evaluator rated performance more in the middle of the scale (i.e., 2 and 3 ratings) and fewer extreme ratings (1 and 4 ratings) than the referent.

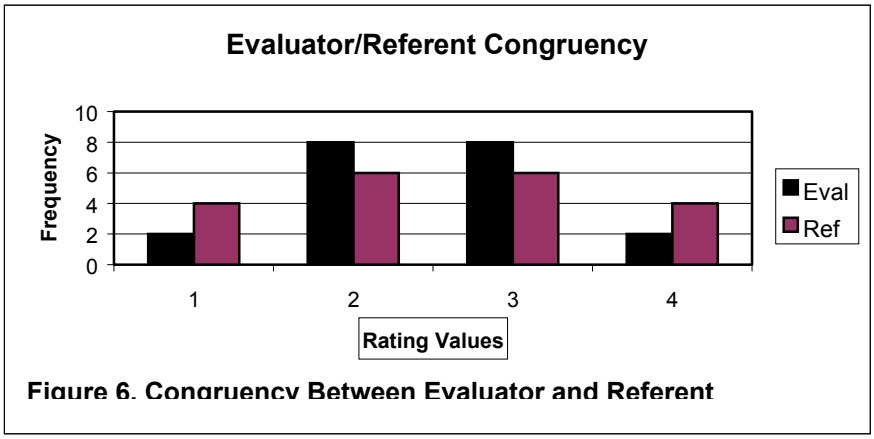

Figure 6. Congruency Between Evaluator and Referent

The congruency measure can not be truly classified as a sensitivity or accuracy type of measure. High congruency neither ensures high sensitivity nor high accuracy for the simply reason that congruency is only concerned with the frequency in which various ratings are used, not if a specific rating is appropriately high or low. On the other hand, if MAD is near 0.0, congruency would necessarily be very high so there would be little need to look at congruency.

In summary, congruency can be viewed as a useful diagnostic when MAD and RRR are lower than what is desired.

Validity

Definition

Validity is concerned with the question of whether an instrument measures what it is assumed to measure. In the case of many physical properties (e.g., weight, color, etc.) there is little concern that our scale is truly measuring weight or that our tape measure is truly measuring length. However, in the case of many behavioral or performance measures there is often a great deal of concern regarding validity. A classic example of this is the concern regarding intelligence tests and whether they are truly measuring intelligence. Albeit to a lesser degree, the same concern can be expressed regarding various measures of pilot performance. For example, line check performance is assumed to measure how a crew operates an aircraft under actual flying conditions. However, it might be found that aircrews are on their best behavior during a line check evaluation and the moment they are no longer being monitored their technical and management performance deteriorates. If this were to happen, the line check data would not be a valid measure of how the crew flies the aircraft under normal every day conditions.

Clearly, if our measures are not measuring what they were designed to measure, we do not have quality data. In addition, validity requires reliable and accurate measurement. If a measuring instrument is insensitive or inaccurate it is severely limited in its ability to measure any property of the world. In this regard, validity is the final challenge to achieving quality data. As will soon become apparent, demonstrating validity in our performance measures is an extremely complex and nontrivial problem. To begin, our measures must first be reliable and actuate. Low reliability or accuracy implies poor validity, but high reliability and accuracy does not imply high validity.

There are three basic types of validity; content validity, predictive validity, and construct validity and while it would be desirable to demonstrate that our measures had all three types of validity, the case for validity can be made by demonstrating any one of the three types. Before proceeding with our discussion of the three types of validity the reader must be forewarned that our discussion of validity will take a far less definitive tone than was taken with the sections on reliability and accuracy. In the case of validity we unfortunately do not have a prescribed set of methods that will ensure validity. Rather we will suggest some strategies that will possibly

improve the validity of our air crew performance measures. In sum, validity is an ongoing process; a goal that we move towards, but never establish in a clearly definitive manner.

Content Validity

Content validity refers to the extent to which the contents of the measuring instrument or test corresponds to what you are attempting to measure. Let us assume for the moment that in the case of crew performance we are attempting to measure how safely and efficiently a pilot operates his/her aircraft on a regular basis. This being the case, we could argue that LOEs, maneuver validations, or line check performance measures are valid to the degree that their content is similar to the content of flying the aircraft on a daily basis. Or we might want to make the case that our ultimate goal is to reduce the incidence of various types of incidents for which there are well documented records. Thus, if we design event sets and LOEs to simulate these incidents we may again argue that our performance measure has content validity.

From our discussion of content validity it can be seen that the measurement of content validity is often quite subjective, although in some instances it may be possible to conduct a detailed content analysis and quantitatively estimate the proportion of relevant content that is sampled by the measuring instrument.

Predictive Validity

Predictive validity is simply the correlation between the measuring instrument and some external criterion that represents what you are attempting to measure. For example, assume we had a test that was purported to measure stockbroker's skill at picking stocks. The predictive validity of this test would be established by simply correlating each brokers test score with how well his stocks performed over some time interval. If we find that there is a high correlation between test scores and stock performance, then the test has demonstrated high predictive validity. Here it can be clearly seen that if our test were unreliable or inaccurate it would limit the magnitude of its correlation with the external criterion and thereby set limits on predictive validity.

There is a lot to be said for predictive validity. It is relatively simple, direct, and quantifiable. However, it does require the identification of an external criterion and that is the rub in using predictive validity in the context of crew performance. Again, assume that our ultimate concern is the safe and efficient operation of the aircraft. What does this suggest as an external criterion? The most obvious external criterion would be line- check data, which might

be assumed to reflect the everyday level of performance of a crew. Unfortunately, there are at least two potential problems in using line-check data. First, the presence of the evaluator in the flight cabin may affect crew performance and invalidate it as representative of everyday performance.

Second, it could not be used as an external criterion for LOE performance, because LOE performance is more concerned with abnormal flight conditions, whereas the vast majority of line-check rides will only sample performance under normal flight conditions. There is no assurance that a pilot's performance in an LOE would necessarily correlate that highly with her performance under normal conditions.

In summary, while both LOE and line-check performance may be valid measures, they would not necessarily be expected to correlate very highly with one another. For this reason, line-check performance may not serve as a good external criterion for LOE performance. More generally, it is quite possible that there is no single external criterion that may be used to validate LOE performance. It is with this possibility in mind that we propose what may be referred to as a multi-pronged assessment of LOE validity. The logic of this approach is that while there is no single external criterion by which to validate LOE performance, there are a variety of criteria that may be moderately correlated with LOE performance. We suggest that a pattern of moderate positive correlations may function to establish the construct validate LOE performance.

Construct Validity: A Multi-pronged Assessment of LOE Validity

As was suggested earlier, in complex and diverse domains such as aircrew performance, there is no explicit set of methods that ensure the validity of our measures. Rather, there are some strategies that are likely to improve their validity. This becomes apparent in our discussion of construct validity as an approach to improving validity. The overall strategy suggested by a construct validity approach is to find a pattern of relationships that is consistent with our general theory of what underlies the safe, competent and skillful operation of the aircraft.

For example, we might hypothesize the following four general factors as underlying the skillful operation of the aircraft: 1) Social interpersonal skills; 2) Cognitive skills; 3) Technical declarative types of knowledge; and 4) Technical psychomotor and perceptual types of procedural skills. We could then proceed to look to different types of measuring instruments that assess these various skills and knowledge. LOE performance might be hypothesized to depend

most heavily on social-interpersonal and cognitive skills, whereas first-look maneuvers may depend more heavily on technical knowledge and psychomotor skills. Within LOE assessment we could further analyze performance on the basis of event sets that are more dependent on social-interpersonal versus those that appear to be more dependent on cognitive skills. Similarly, critical maneuvers could be further analyzed using systematic task analyses and judgments of subject matter experts into those that are more dependent on technical knowledge versus psychomotor skills.

At this point we could begin to look at pilots' performance on these various tasks to determine if the patterns of correlations are consistent with our hypothesized model. For example, we should expect to find that performance on event sets measuring social-interpersonal should be more highly intercorrelated with one another, than they are with event sets that were judged to be more related to cognitive skills (e.g., decision making). At the same time, performance on these two types of event sets should be more highly correlated with one another than they are with performance on critical maneuvers. Within critical maneuvers performance we should expect to see those maneuvers that are more knowledge dependent correlate more with event sets that are cognitively based than the more social-interpersonal based event sets.

Similar types of analyses can be conducted that look at the relationship between training and performance. For example, if LOE performance reveals a deficit in situational awareness we would then want to strengthen training in this area. If we later see an improvement in performance in the context of LOE evaluations we have validated a relationship between the assessment and training of situational awareness. The content of the curriculum on situational awareness, in effect, tells us in part what is being measured by those specific event sets.

Finally, there is a relatively new source of data, Flight Operations Quality Assurance Program (FOQA) that has the potential of contributing significantly to the construct validity of our current performance measures. FOQA involves a technology that allows for the continuous recording of many physical parameters related to flight information. Using various algorithms it is possible to extract composites of flight data that meaningfully reflect technical skills that are related to qualification standards. When these data are de-identified, it would remain possible to related the incidence of various exceedances to fleet aggregated LOE, Maneuvers, and Line Check data. Part of the validity of LOE assessments of management skills rests on the assumption that poor CRM is eventually manifested in diminished technical skills. If this

assumption is valid, it should be empirically supported by a relationship between fleet aggregated FOQA and LOE data.

In summary, when taking a construct validity approach no single correlation is critical. Rather, it is the general pattern of correlations and the degree to which they are consistent with our model of what our measures are assessing. The viability of this approach rests on our ability to analyze and classify data from each of many different sources (e.g., LOE, maneuvers validation, first-look, check-rides, FOQA, etc.). The model for doing this is contained in the links between the Audit Proficiency Database and the Performance Proficiency Database. This structure makes explicit the kinds of interrelationships we should expect to find in our correlational analyses.

Quality Aircrew Performance Data

Now that we have a better understanding of what quality data entails, we can shift our attention to the process of implementing quality data in the assessment of aircrew performance. We begin with a discussion of why it is so important for a carrier to have quality data. Next, we turn our attention to the various factors that can influence reliability and validity when assessing aircrew performance and what can be done to improve the quality of the assessment process.

Importance of Quality Data

Having discussed the formal properties of quality data it is important to understand why it is so important for an airline to have a reliable, accurate and valid means of assessing pilot performance. Every airline requires quality data for three basic reasons; detecting *what* is changing; detecting *when* it is changing; and detecting *who* is changing.

What is Changing. Perhaps the most basic reason for requiring quality data stems from the close relationship between assessment and training. Quite simply, the quality of training can be no better than the quality of the data used to assess the training. This relationship between training and assessment is the fundamental core of AQP. Under AQP it not sufficient to simply train. It must be demonstrated that the training ensures proficiency and this can only be accomplished with quality assessments that tell us precisely what aspects of the curriculum and training are working and what components are not working. Only then can training be focused where it is most needed.

When there is Change. With the development of quality measures of crew performance it is possible for carriers to do a better job of tracking changes in performance over time. Assume that a fleet is tracking the mean LOE performance of its
crews over an extended period of time. Figure 6 presents two plots of how these data might look for reliable and unreliable measures of performance. Although the trends in these two panels are actually identical, it is far more difficult to detect the downward trend with the higher variability plot. This is a simple illustration of how reliability influences variability, which, in turn degrades the precision of decision making. More reliable performance measures allow a carrier to more quickly and accurately detect trends and take corrective action.

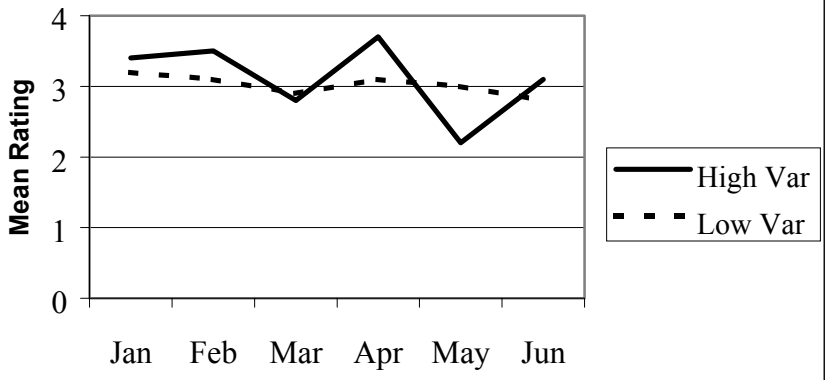

Figure 6. Control Chart for High and Low Variability Data

Who is Changing. Finally, while the primary goal of assessment it to improve training, quality data also serves as the basis for sound and rational personnel decisions. In any organization employing a large number of individuals, personnel decisions must be made. Given these decisions are made, it is best if they are made on the basis of quality data. To maintain a highly competent and dedicated group of employees it is essential that the employees recognize that management appreciates their efforts to excel at their job. For this to happen, management must be able to distinguish who is performing at a superior level and who is performing at a less than acceptable level. This requires that management be able to assess knowledge and performance in a reliable, accurate and valid manner. If the measures are unreliable, inaccurate

and invalid there is no fair basis for advancement and morale problems will soon develop.

Three Elements of Assessing Evaluator Performance

The assessment of aircrew performance typically involves an evaluator, a measuring instrument (e.g., an LOE grade sheet), and a specific set of conditions under which the evaluator and the measuring instrument are themselves evaluated (e.g., an evaluator calibration session). In this section we consider how each of these factors can influence the reliability and validity of the performance data and what can be done to improve the quality of these data.

Evaluator Reliability: Sensitivity and Accuracy

The evaluator plays a central role in the quality of LOE, First Look, Maneuvers Validation, and Line Check data. Therefore, it is of the utmost importance that we are able to assess the reliability of every evaluator's judgments. When an evaluator's performance is below standards, the assessment should tell us where training needs to be focused.

There are only two types of errors an evaluator can make. An evaluator can be insensitive or inaccurate. Sensitivity is measured by RRR and accuracy is measured by MAD. Given that each of these two types of problems may either be present or not, there are four possible diagnostic categories (See Table 4). Each of the four categories in this diagnostic matrix has clear implications for training. If both RRR and MAD are good (high RRR and small MAD), no additional training is required at this time. If RRR is bad and MAD is good, the evaluator has a sensitivity problem and needs training on discriminating different levels of performance. As noted earlier, it is logically impossible for RRR to be extremely low if MAD is extremely small. If RRR is good and MAD is bad the evaluator has an accuracy problem and it suggests that training should be focused on the use of the grading scale. Examination of the evaluator's distribution of grades compared to the referent will indicate if the grading is too generous or too harsh. This kind of feedback, which is provided as part of the calibration session, may be sufficient to correct a simple accuracy problem.

Finally, if an evaluator is weak on both RRR and MAD, a look at the evaluator's mean rating compared to the mean for the referent will reveal if the large MAD is caused by the low RRR. If the mean rating is fairly close to the referent mean, it suggests that the large MAD is driven by the rater's insensitivity (e.g., an RRR of –1.0 would necessarily result in a large MAD). However, if the evaluator's mean rating is substantially above or below the mean for the

referent it suggests that the evaluator truly has both a sensitivity problem and an accuracy problem.

Table 4. Evaluator Diagnostic Matrix.

		RRR: Good	RRR: Bad
MAD	Good	No training necessary	Performance Sensitivity Training
MAD	Bad	Scale Accuracy Training	Performance Sensitivity Then Scale Accuracy

Evaluator Validity

Our concern here is that the evaluators' judgments are based on the appropriate qualification standards when grading aircrew performance. When grading LOE, First Look, Maneuvers Validation, or Line Check performance, an evaluator must know what qualification standards are relevant for each phase of flight and each situation (e.g., event set) within a phase of flight. Without this knowledge an evaluator cannot validly grade performance. To the extent that the standards for management skills are any less explicit than the standards for technical skills, we might be more likely to encounter a validity problem in evaluators' judgments of management skills. Knowledge of qualification standards may be assessed directly with a paper and pencil type of test. Ensuring that evaluators are grading on the basis of qualification standards will have the most direct positive effect on content validity. However, as the content validity of the evaluators' judgments improves, we should also expect to see an improvement in predictive and construct validity.

Earlier in this chapter we discussed how poor reliability lowers validity. Here is a situation where poor validity could lower reliability. If evaluators have a poor understanding of qualification standards, not only are they more likely to be grading on the wrong basis, they are also less likely to be in high agreement with one another, resulting in a lower RRR.

Instrument Reliability: Sensitivity and Accuracy

Here we are concerned with the influence of the measuring instrument on the reliability of our measurements of aircrew performance. For example, a poorly designed LOE grade sheet can adversely affect both sensitivity and accuracy. It has been shown that any vagueness in the phrasing of the performance indicators on the grade sheet dramatically lowers RRR and IRR. Below we discuss how relatively minor changes to clarify the wording of a performance indicator can greatly increased evaluator agreement.

Turning to accuracy, the grade sheet should provide evaluators with clear instructions on the appropriate use of the grading scale. These instructions should appear on the grade sheet and in addition it is recommended that evaluators be given more elaborate instructions on the use of the grading scale before beginning a calibration. If, for example, a 4-point grading scale is being used, it is helpful if the evaluators are provided with several examples of what constitutes a 1, 2, 3, or 4 level of performance.

Instrument Validity

Just as our concern with evaluator validity was in ensuring that evaluators were grading on the basis of qualification standards, the same holds true for instrument validity. The items comprising the measuring instrument should be as closely related to the qualification standards as possible to ensure content validity. For example, in the case of assessing LOE performance, it should be possible to relate every performance indicator to a knowledge or skill in the qualification standards. Once again, as content validity improves we would expect to see a corresponding improvement in predictive and construct validity.

Despite the difficulties in assessing predictive and construct validity, it should be possible to determine how changes in a performance measure affect correlations with various constructs. Continual improvements in content validity should eventually result in gradual improvements in a measuring instrument's construct validity.

Situation Reliability

Before proceeding with our discussion of situation reliability the reader should be reminded that we are referring to the conditions under which we calibrate an evaluator and the measuring instrument (i.e., the calibration session). Once again, while we are ultimately interested in the quality of actual LOE, Maneuvers Validation and Line Check data, it is often

impossible to obtain reliability estimates in these situations (e.g., it is necessary to have every evaluator view the exact same performance to compute RRR or IRR). Thus, we resort to videos and calibration sessions to provide reliability estimates. There are a number of factors surrounding a calibration session that could lower reliability. Here we consider three types of situational factors; viewing conditions; video quality; and instructions.

Viewing Conditions. The influence of the viewing conditions on reliability can be easily illustrated in the context of an LOE calibration session where a large number of evaluators are in a single room viewing an LOE video. Under these conditions numerous potential error sources are introduced. The viewing and listening conditions in the room will vary depending where an evaluator is seated relative to the screen and the speaker. If evaluators are talking to one another during the showing of the LOE video this will introduce another source of error that is likely to lower reliability.

There are some fairly obvious precautions that can reduce most of these sources of error, however, it is important to keep in mind that we are attempting to estimate reliability as it occurs in the operational situation. For example, with an LOE calibration session we are attempting to estimate evaluator reliability as it occurs in a full flight simulator. The viewing conditions in a calibration session differ in many ways from what happens in the simulator. The level of workload in the simulator is likely to be far greater than is present in a calibration session. Thus, if high workload functions to lower reliability it is possible that we are overestimating evaluator reliability in our calibration sessions. On the other hand, it is possible that some of the technical information relevant to arriving at a judgment is more available in the simulator than on a video of a crew flying the aircraft. The point is that whenever possible we should strive to make the viewing conditions in the calibration session as similar as possible to what the evaluator encounters in a full flight simulator.

Video Quality. Of all the factors that we consider, the video itself, both in terms of its content and the quality of the audio and visual signal, may have the single greatest impact on reliability.

All the information necessary to grade each item on a grade sheet must be clearly presented in the video. Often this may include conversations among crewmembers, or it may involve technical information requiring a clear view of the relevant instrument readings.

Evaluators should not be expected to grade the omission of some action or decision, unless there is an explicit context indicating where the event should have occurred.

Once the video is developed in concert with the grade sheet they must be examined as a unit by a group of experienced evaluators. This involves having the group of experts view and grade the video, ensuring that there is an objective basis for grading each item. In every instance the experts must agree as to the key information in the video, how the behavior was consistent or inconsistent with qualification standards, and what the *referent grade* should be on each item. If there is a lack of strong consensus, either the video or the grade sheet needs to be modified to ensure there is high consensus among experts.

Finally, when we are using a video to simulate some performance situation (e.g., a crew flying an LOE), the evaluator's familiarity with flight scenario shown in the video should be comparable to his or her familiarity with LOEs occurring in the simulator. Under most conditions evalutors will be highly familiar with the LOE they are running in the simulator. If this is the case, then care should be taken that evaluators are also highly familiar with the LOE shown in the video. It is suggested that approximately one week before the calibration session is scheduled to take place, the evaluators be given a copy of the grade sheet to allow them to become familiar with the event sets, the performance indicators, how and where ratings are entered, etc.

Instructions. Here, we refer to instructions in the most general sense of preparing the evaluators for the calibration session. Setting the context for the task and what the evaluators are expected to do is essential to obtaining quality data. It is also essential that the evaluators appreciate why they are participating in this process and why it is so important to the mission of the carrier. This is part of the process of facilitating evaluator buy-in with respect to calibration sessions and collecting quality data. As much as possible, evaluators and their supervisors should be brought into the development of all phases of the calibration sessions (i.e., the grade sheet and video).

Situation Validity

Many of the same situational factors that were discussed above as influencing reliability also may be expected to affect validity. Our central concern with respect to validity is that the evaluators behave as closely as possible to how they would behave in the situation that we want to generalize to. Thus, the situation validity of a calibration session depends on how closely it

approximates the situation that exists in the simulator. Much of this may depend on evaluator motivation and "buy-in". If the evaluator perceives the situation as realistic and is motivated to do his or her best job, we are a long way toward achieving situation validity. Again, the same types of factors that influenced situation reliability are relevant here. If the viewing conditions more resemble a party atmosphere than a serious evaluation, we know we have a validity problem. If the video is unrealistic in any regard it will likely diminish motivation and buy-in. Finally, it is of the utmost importance that the supervisor running the calibration session sets the appropriate tone when delivering the instructions. The evaluators have to be made aware of the importance of quality data and the role of the calibration session in achieving quality data. Only then can we expect to get the level of motivation and buy-in that is necessary to ensure valid data.

LOE Calibration

In this section we discuss the three phases of conducting an LOE calibration session with a group of evaluators. The three phases are data collection, data analysis, and feedback. In our discussion of these three phases it is assumed that they are completed on a group of evaluators within a single day. The data collection phase is usually completed in the morning, allowing two to three hours to analyze the data and generate reports, and then concludes with the feedback phase in the early afternoon.

Data Collection Phase

In this phase the evaluators are asked to evaluate a video of a crew flying an LOE. To facilitate buy-in an evaluator supervisor who they know and respect should conduct this phase of the calibration session. The session begins with the supervisor giving a general overview of the sequence of events in the calibration session. The grade sheets are then distributed and the supervisor walks the evaluators through all aspects of the grade sheet. It is at this time that any potential ambiguities in the phrasing of an item are clarified. In addition, any uncertainty regarding the grading scale for event sets and the performance indicators is clarified with concrete examples. Despite the fact that all of the evaluators should be experienced with the grading scale, it is necessary to re-affirm the proper use of the scales for grading event sets and performance indicators.

When the evaluators are ready to proceed with the grading of the video the supervisor sets the context of the flight scenario that is contained in the video, concluding with a heads-up

on what will be occurring in the first event set they will view and grade. Evaluators are instructed that they may either grade the items as they occur in the video or wait until the end of the event set to enter their grades. If they wait, they are reminded to make notes during the video. Finally, they are instructed to focus their attention on the video, to make their ratings independently and to not engage in conversation with their neighbor.

At the end of each event set they are given time to complete entering their grades and the supervisor then sets the context for the next event set. After viewing and grading all of the event sets, they are reminded to be certain their PIN is clearly written in the appropriate location on the grade sheet and the grade sheets are collected.

Data Analysis Phase

In this phase of calibration the LOE data are transcribed to a computer file for statistical analysis and report generation. To facilitate the speed and accuracy of this process, we have developed a PC ACCESS based software calibration package that expedites the data entry and automatically analyzes the data to generate the necessary individual and group statistics.

Evaluators' ratings are entered directly on the computer screen which displays a form closely resembling the gradesheet. Once these data are entered for all of the evaluators participating in the calibration session, the group and individual statistics are automatically computed. Table 5 is an example of individual report that is generated for each evaluator, showing how he/she performed relative to the referent on each performance indicator. At the bottom of this report the evaluator's average rating, MAD, RRR and IRR performance is summarized.

Table 5. Individual Summary Information

Name: obs1 *PCA/APD: PCA*

ID: 1 *Fleet:* 767

Event Set Number	*Type*	*ItemText*	*My Score*	*Qualification Standard*
1	M	(DM) Complies with Standard Policy for Takeoff and Go/No Go Decision	2	3
1	M	(SA) Commands a maneuvering airspeed consistent with aircraft configuration	3	3
1	M	(CC) Keeps PNF informed of intentions.	1	3
1	T	Maintains effective aircraft control throughout the takeoff event.	3	3
1	T	Accomplishes After Takeoff checklists IAW the POM.	2	2
1	T	Accomplishes Hydraulic Abnormal checklists IAW the POM	3	3
2	M	(CC) Completes Approach briefings (NATS)	2	3
2	M	(CC) Coordinates use of Autopilot Flight Director System	1	1
2	M	(PL) Proactively plans to remain ahead of aircraft/situation	1	1
2	M	(WM) Distributes workload effectively	2	1
2	T	Complies with Standard Policy for checklists.	1	1
2	T	Performs non-precision approach IAW POM, Maneuvers section	3	1
2	T	Performs missed approach procedures IAW POM, Maneuvers section	1	2
3	M	(CM) Communicates intentions with ATC after engine failure.	1	3
3	M	(WM) Prioritizes tasks of flying the departure and completing the abnormal.	2	2
3	M	(CC) Calls for appropriate checklists.	1	3
3	T	Maintains effective aircraft control throughout the takeoff event.	3	1
3	T	Accomplishes after takeoff checklists IAW the POW.	1	3
3	T	Accomplishes engine failure after V1 procedures and maneuvers IAW the POM	3	2

Average Scores

Individual: 1.89
Referent: 2.16
All Participants: 1.85

Mean Absolute Difference from Referent

Management: 1.00
Technical: 0.56
Overall: 0.79

MyScores Correlated with Qualification Standard

Management: 0.509
Technical: 0.795
Overall: 0.305

My Scores Correlated with Other Evaluators

Management: 0.075
Technical: 0.404
Overall: 0.435

Table 6. Rank Order of Items by Mean Absolute Difference

Event Set	Type	Number	ItemText	MAD
2	M	3	(PL) Proactively plans to remain ahead of aircraft/situation	0.00
1	M	3	Accomplishes Hydraulic Abnormal checklists IAW the POM	0.00
2	M	2	(CC) Coordinates use of Autopilot Flight Director System	0.20
2	T	4	(WM) Distributes workload effectively	0.20
3	M	2	Accomplishes after takeoff checklists IAW the POW.	0.40
2	T	1	Complies with Standard Policy for checklists.	0.40
3	M	2	(WM) Prioritizes tasks of flying the departure and completing the abnormal.	0.60
2	M	2	Performs non-precision approach IAW POM, Maneuvers section	0.60
1	T	2	Accomplishes After Takeoff checklists IAW the POM.	0.60
1	M	1	Maintains effective aircraft control throughout the takeoff event	0.60
3	T	3	Accomplishes engine failure after V1 procedures and maneuvers IAW the POM	0.80
2	T	3	Performs missed approach procedures IAW POM, Maneuvers section	1.00
1	T	1	(DM) Complies with Standard Policy for Takeoff and Go/No Go Decision	1.00
1	T	2	(SA) Com mands a maneuvering airspeed consistent with aircraft configuration	1.00
3	T	1	Maintains effective aircraft control throughout the takeoff event.	1.20
1	M	3	(CC) Keeps PNF informed of intentions.	1.20
3	M	3	(CC) Calls for appropriate checklists.	1.40
3	T	1	(CM) Communicates intentions with ATC after engine failure.	1.60
2	M	1	(CC) Completes Approach briefings (NATS)	1.80

Table 7. Individual Summary, EventSet

Name: obs1 *PCA/APD ID: PCA*
ID: 1

Event Set	My Score	Qualification Std	Group Avg.
1	3	3	3.200
2	2	1	3.000
3	3	1	2.600

My Average Event Set Score: 2.67
Qualification Standard Average Event Set 1.67
MAD between My Scores and Qualification Std: 1.00
MAD Across all Individuals: 1.27

Table 8. Performance Indicator Summary

Average of Correlations with Qualification Std

Management:	*0.80*
Technical:	*0.83*
Overall:	*0.81*

Average Correlations with Other Evaluators

Management:	*0.76*
Technical:	*0.79*
Overall:	*0.78*

Mean Absolute Difference with Referent

Management	*0.12*
Technical	*0.10*
Overall	*0.11*

Thursday, January 22, 1998 *Page 1 of 1*

Table 6 shows an example of a report that rank-orders the items from lowest to highest MAD score. The report shown in Table 7 summarizes event set ratings for the group and the individual. Finally, Table 8 is an example of the reported generated on group RRR, IRR, and MAD statistics. Broken out in terms of management and technical performance indicators. With this software the data entry and analysis can usually be completed for 50 to 60 grade sheets within two hours.

Feedback Phase

Conducting the feedback phase of the calibration session may require a two-person team, comprised of an individual who is familiar with all of the statistical procedures used to analyze the data and an evaluator trainer who is expert in the qualification standards and the grading scale. Because this is where the most of the important training occurs it important to allow sufficient time to explain all of the results and address the evaluators' numerous questions and comments.

The feedback phase begins with the distribution of the group and individual data summary sheets to the evaluators. The individualized reports allow each evaluator to see his or her performance on all of the measures relative to the group average and the referent. In addition, each evaluator can see how his ratings compared to the group and referent on each event set and performance indicator on the grade sheet. After the group and individual reports have been returned to the evaluators the discussion leader first provides a brief overview of what information is contained in each of the tables and reminds them that the primary purpose of the calibration is training and fine tuning of the instruments.

While there is no particular order that needs to be followed in discussing the results with the evaluators, it may help to relax the group by beginning with a discussion of performance on the highest agreement items in the LOE grade sheet (see Table 6). To the right of each performance indicator is the MAD for that item. The items have been rank-ordered with those having the highest agreement at the top of the table. From Table 6 it can be seen that MAD is equal to 0.00 for the highest agreement items, indicating that all of the evaluators gave that item the same rating. In addition to praising the evaluators for their performance on these items, these results also clearly establish that it is, indeed, possible to attain perfect agreement.

The discussion next turns to those items for which there was the highest disagreement (see Table 6). When turning to the low agreement items it is important to emphasize that the goal here is to "fix the bad items". For these items it is important to make every effort to determine the source of this variation. Toward this end the discussion leader should encourage the evaluators to communicate the basis for their ratings on each of these high disagreement items. The goal of the discussion leader is to discover what produced the high levels of disagreement on each of these items. This requires the active participation of the evaluators and the discussion leader's explanation of why they graded the item so much lower or higher than the referent and

most of the other evaluators. Some evaluators may find this is a threatening situation. The discussion leader can reduce some of this tension by creating an atmosphere where the evaluators are working as a team to improve the quality of the work sheet and the video. Once the sources of disagreement on a particular item are understood it is often a relatively simple matter of rewording or elaborating the description of the performance indicator. For example, in a previous calibration session there was found to be high disagreement on the performance indicator "Engine-out missed approach procedures". Rewording this item to "Performs engine-out precision approach procedures and maneuvers IAW POM" reduced the disagreement by 63% on a second calibration session. Information on how best to word an item can often be obtained from evaluators comments elicited during a calibration session.

It is also important to look for patterns in the analysis of item performance. If, for example, it is discovered that there are a disproportionate number of high disagreement items related to decision making, the discussion leader can focus his or her questions on this area during the feedback phase. As part of this discussion it may be discovered that when evaluators are asked to rate "crew exercises good decision making", they are uncertain whether to rate the item on the appropriateness of the crew's final decision or on the basis of the process by which the decision was made. This suggests that the problem is not specific to particular items, but is more generic and therefore may not require rewording all the items related to decision making.

The final topics of discussion are the group and individual statistics. Much of the discussion here will be directed towards explaining what the RRR, IRR, and MAD statistics are measuring. Again, the focus of the discussion should be in terms of implications for training. Noting that a low RRR calls for sensitivity training and a high MAD calls for training on the grading scale. It may be useful at this juncture to present Table 4, which shows the different combinations of good and poor performance on RRR and MAD. The evaluators could then be walked through the four matrices of Table 4, explaining what each of the cells indicates.

Because each evaluator has the data summary sheet showing his/her performance in comparison to the referent and group performance statistics it is unnecessary to discuss any individual's performance. Each evaluator will be able to clearly see how he/she did relative to the referent and other evaluators. It is only necessary to discuss patterns of errors and what they may indicate (e.g., commenting that the mean rating on this item was a 3.2, whereas the referent rating was a 2, suggesting that many of their ratings on this item was too high). Some general

bench marks may also be helpful in letting an evaluator know when he is being too unreliable or inaccurate (e.g., if your rating was higher than the referents on 16 or more of the 20 items you very well may be grading too leniently).

Evaluating Calibration Training

To assess whether calibration training is improving the quality of assessment data it is necessary to generate control charts on evaluators' RRR and MAD for each fleet over extended periods of time. Figure 7 shows a control chart plotting RRR performance over one-month intervals. A certain degree of variability in these measures will occur simply from sampling error (i.e., the sample of evaluators will change across calibration sessions). However with the appropriate statistical methods it is possible to set upper and lower confidence intervals that will distinguish real changes from random sampling variations (See Figure 7). If the variation exceeds either the upper or lower boundary it indicates that a real change was observed.

In Figure 7 it can be seen that within the month to month variation there is a gradual trend

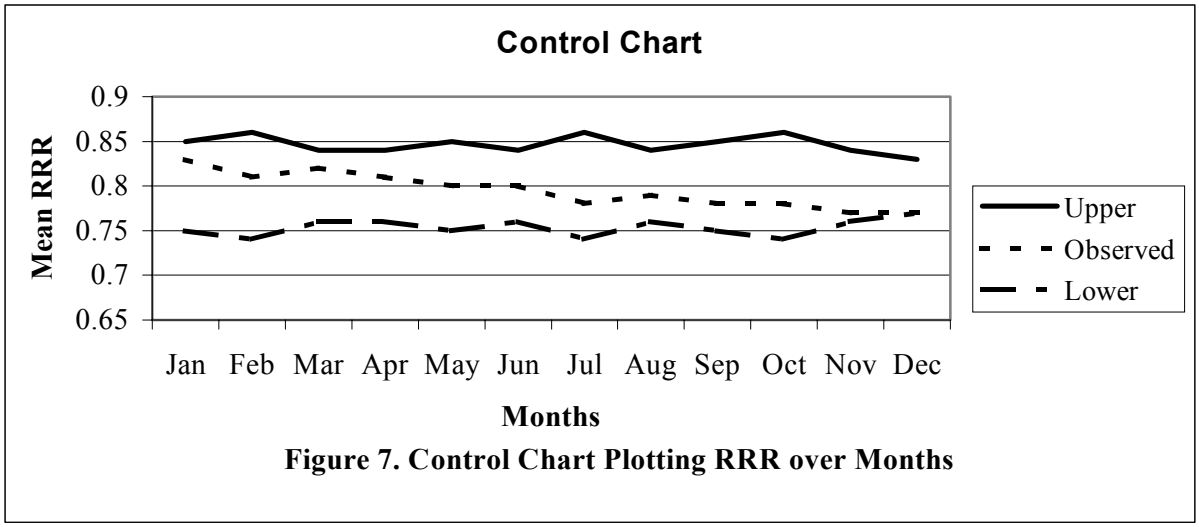

Figure 7. Control Chart Plotting RRR over Months

towards lower RRR values. Again, with the appropriate statistical analysis (e.g., a Regression analysis) it is possible to determine whether this trend is statistically significant or not. If the calibration training is truly having a positive effect the control charts should eventually begin to show improvements in RRR, and MAD.

Two major questions regarding the effects of calibration training are its longevity and generality. We would certainly like to believe that training has relatively long lasting effects and that training on one set of event sets would generalize to new event sets. Of course, this is an empirical question that can only be answered with the appropriate data. Unfortunately, at this

time there are no reliable data on either the longevity or the generality of evaluator calibration training. This problem can be addressed by establishing an Evaluator Performance Proficiency Database. As we begin to collect data from repeated calibration sessions it will be possible to generate control charts that plot evaluators' performance across months. If these control charts fail to show any improvement in RRR or MAD after several months of calibration training it would be necessary to conduct a more detailed analysis of these data. First we would want to determine whether there is any improvements when evaluators are retested on the same event sets they were previously trained on. Ideally, we would initially want to look at these effects over relatively short intervals (e.g., one-month intervals) and gradually extend them to estimate the duration of calibration training effects. If there are no benefits after a one-month interval it would probably be necessary to re-evaluate the nature of the calibration training sessions.

Once it has been shown that the training has reasonable longevity, we can begin to look for transfer effects (i.e., does training on event set A facilitate performance on evaluating event set B). The central question here concerns how diverse the training must be (i.e., how many different event sets are calibrated), before we begin to see beneficial transfer effects to new event sets. To address this question requires that the Evaluator Performance Proficiency Database be structured to allow the assessment of transfer effects as a function of the number of previously calibrated event sets.

EXTENDING QUALITY DATA METHODS TO ALL PERFORMANCE MEASURES

Although the methods for achieving quality data that are discussed in this chapter are intended to apply to all pilot performance measures, the examples have most often used LOE performance indicator ratings. In this section we discuss specific issues that arise in the application of these general principles to LOE event set ratings and maneuver validation ratings.

Event Set Ratings.

. Two specific issues arise in regard to assigning a global rating to an event set. First, because there are many fewer event sets than performance indicators that are rated in an LOE, there arises a potential sample size problem (e.g., there may be too few event sets to obtain an accurate estimate of reliability). A second potential problem arises from the fact that event set ratings are more global and subjective than performance indicator ratings and as a consequence it may be more difficult to identify specific sources of disagreement in calibration training

Sample Size Problem. Earlier, we discussed how the stability or meaningfulness of RRR statistics is dependent on the number of items they are computed on and how this could present a problem with some calibration videos. While it is somewhat arbitrary to set an absolute minimum number of event sets for computing RRR, it would be conservative to say that as the number of events sets goes below eight there would be little practical value in computing correlational statistics. It would be desirable to have 20 or more event sets to have sufficient power to detect real differences among evaluators. Given that it is unlikely that more than eight event set ratings will be collected from an evaluator on a single calibration it is recommend that evaluators be presented only the descriptive statistics shown in Table 7.

The reader should be reminded that the reliability of the group MAD statistic is dependent on the number of evaluators, not the number of event sets. Therefore, if there are a reasonable number of evaluators (e.g., in the range of 20 or more) we can proceed to compute a MAD for each event set. This tells us in an absolute manner, the level of disagreement we are seeing in the rating of event sets and which event sets are producing the highest level of disagreement.

Subjectivity Problem. Whereas ratings on performance indicators refer to relatively specific well defined behaviors, event set ratings refer to a rather extensive and diverse set of actions occurring over several minutes. The rating of an event set is, therefore, a more subjective judgment and this raises the possibility that it may be more difficult to calibrate**.** However, the subjectivity of grading an event set is somewhat lessened if observed behaviors for that event set are also rated. The rating of the performance indicators significantly constrains the rating of the event set.. If a crew fails to perform most or all of the performance indicators listed on the work sheet for an event set it would be difficult to justify an overall rating of "acceptable" or better on that event set.

For the most part we would expect the ratings of events sets and performance indicators to covary. However, because the set of performance indicators listed under an event set is not an exhaustive listing of all the possible important behaviors that could occur in this situation, there is obviously room for disagreements to occur. It is quite possible that a pilot/crew performs all of the performance indicators satisfactorily, but makes a critical mistake that is not covered by a performance indicator. This is one of the reasons that the global rating of the event set is so important and why they will not always agree with the ratings on the specific performance indicators. However, the point to be made here is that when there are disagreements between global and specific ratings, it should be possible to make the basis for this disagreement explicit as part of the discussion with evaluators during the feedback phase of a calibration session.

Maneuver Validations.

The final area of application we discuss concerns the assessment of maneuver validations. Carriers may identified a set of critical maneuvers (e.g., 10) that are used in full-flight simulators to assess pilots' technical skills. In First Look evaluation each pilot is evaluated without any prebriefing on some number of these critical maneuvers. To assess the quality of First Look data it is necessary to develop a video of a crew flying the different maneuvers. Ideally, we would want to have a video of each critical maneuver and then at least two levels of performance for each maneuver. However, it is not necessary to have this entire library of videos before a carrier begins a calibration program. As was the case with the LOE videos it is necessary to have a group of experts validate the videos and arrive at a referent grade for each one.

The Maneuvers calibration session would proceed in a similar manner as an LOE calibration session. A group of evaluators would be shown a video of a crew executing a set of maneuvers. At the end of each maneuver, each evaluator would independently grade the performance on a 4-point scale. Once the ratings had been collected on all maneuvers, RRR and MAD can be computed in the same manner as previously described for performance indicators. The same concerns regarding the reliability of our statistical estimate of RRR that arose with a decreasing number of event sets in an LOE would apply with first look maneuvers. If the number of maneuvers decreased below nine the reliability of our estimation of evaluators' RRR would diminish rapidly.

Conclusions

This concludes our discussion of what it means to collect quality data, why it is important to the well being of the carrier, and what can be done to insure the collection of quality data in all aspects of training and assessment. In reading this document it becomes apparent that the collection of quality data is a multi-step process and the final product is only as good as the weakest link in the chain. The development of measuring instruments, the conditions under which they are administered, etc., are all important. However, none is more important than the evaluator. Quality data can only be attained with the involvement of a dedicated and highly skilled staff of professional and highly trained and calibrated evaluators.

Assessing and Improving Evaluation of Aircrew Performance

Proper evaluation of human performance in the workplace is important to the success of both the individual worker and the employer. Historically, psychologists have devoted much attention to this topic by offering tools for assessing the quality of evaluations and improving the evaluation process (see Arvey & Murphy, 1998, for a recent review). Although the specific issues of performance evaluation may change over time, the general goal of these evaluations remains constant: to seek fair and accurate evaluations of an individual's work performance.

Few employees are scrutinized as closely as airline pilots. There exist well-defined evaluation cycles, specific skills and knowledge to be tested, objective criteria on which to judge performance, tight restrictions on evaluator qualifications, and well-specified criteria of realism for creating an evaluation context. This highly regulated context affords certain advantages for investigating the evaluation process. Further, the quality of aircrew evaluations may exceed that of most other occupations.

Although extensive efforts have occurred at developing and validating assessment instruments for pilots, the bulk of this work has been aimed at selecting pilots rather than on assessing ongoing work performance (Burke, Hobson, & Linksy, 1997; Damos, 1996; Hoermann & Maschke, 1996; Jensen, 1989). Further, selection tests have emphasized personality variables or basic level cognitive and perceptual-motor abilities, neither of which is likely to account for much variance in job performance of highly skilled pilots. Extensive work on aircrew assessment has occurred in the military (e.g., Dwyer, Oser, Salas, & Fowlkes, 1999; Hubbard, Rockway, & Waag, 1989), but these studies appropriately focus on the accomplishment of general mission goals rather than on individual pilot performance. There is also an extensive body of literature on the evaluation of team performance (Baker & Salas, 1996; Salas, Prince, Baker & Shrestha, 1995), which is relevant to pilots performing together as a crew. Our interest, however, is on the evaluation of individual pilots, the primary concern of the commercial aviation community. We use the generic term aircrew to refer both to individual pilots and the flight crew.

Historically, pilots have been expected to demonstrate proficiency on a relatively small set of critical flying maneuvers and knowledge of aircraft systems. Assessing proficiency on these skills and knowledge was viewed as a relatively straightforward task, with pilots assumed to either have or not have proficiency in these areas. Thus, an experienced evaluator would be

able to easily discriminate between the skilled and unskilled. In reality, it is a challenge to obtain reliable judgments of pilot proficiency on even objective technical tasks.

Under newer training and evaluation guidelines, such as those of the Advanced Quality Program (AQP; FAA, 1991), airlines now assess aircrew performance in more complex and ill-defined environments. Pilots are routinely assessed with line oriented evaluations (LOEs), where aircrews fly realistic city-pair scenarios that encompass all phases of flight. Further, these assessments include evaluations of both technical and crew resource management (CRM) skills. Such complex evaluations place new demands on evaluators, who, under AQP, must be trained and assessed in their role as evaluators.

In this paper, we apply some psychometric principles to the evaluation of aircrew performance. A second goal is to offer guidelines for assessing and improving the evaluation of pilot performance, suggestions aimed specifically at designing and implementing training and calibration sessions for evaluators.

Pilot Evaluation Data: Purposes and Sources

We begin by discussing why pilots are evaluated. We distinguish among three primary reasons for evaluating pilots. The first is to decide whether an individual pilot is proficient to fly the line in his current position or is qualified to transition to a new seat position or new aircraft. Second, instructors need to assess their trainees' knowledge and skill in order to offer appropriate feedback and remediation. Although this type of assessment may not occur in the context of a formal evaluation, the quality of pilot training depends on evaluators being capable of making accurate judgments of performance.

Third, aircrew evaluations (should) guide the development and modification of training programs. Under AQP, airlines maintain a pilot proficiency database that houses detailed data on aircrew performance. These data reflect pilot performance on the various skills and knowledge that have been previously defined by job task analyses. These aggregated performance data speak to the strengths and weaknesses of a training program. Under a proficiency-based program such as AQP, a training manager has the flexibility to modify the nature and extent of training when such changes are warranted by empirical performance data. The vast majority of these data are human judgments of pilot performance. Their usefulness to training is directly related to their reliability and validity. It is paramount that training decisions be based on the highest quality evaluation data.

There appear to be two major sources of data from which we can assess the quality of evaluators' ratings of aircrew performance. The first is the pilot proficiency database mentioned above. Later we describe in detail how these data can be used to assess evaluators. A second source of evaluation data, which has arisen primarily as a result of airlines transitioning to AQP, are special evaluator training and calibration sessions. In these sessions, evaluators are asked to observe and grade samples of aircrew performance, typically shown on a videotape or computer screen. Evaluators observe and independently grade the same performance samples. Later we discuss how to develop and implement calibration sessions including what type of statistical feedback to give.

Psychometrics of Aircrew Evaluation

Various statistical procedures that can be used to assess the quality of evaluations. The evaluation of human judgments has a long history in the fields of psychometrics and testing theory (Anastasi, 1988; Nunnally & Bernstein, 1994). Some of our recommendations are based on standard psychometric approaches, and other statistics that we describe were designed specifically for assessing the goodness of evaluations of aircrew performance. Our discussion is divided into two sources of evaluation data discussed above: aircrew performance data and evaluator calibration sessions.

Aircrew Performance Data

We assume that the distribution of task grades given by all evaluators reflects overall fleet performance, along with the airline's established grading policies. If so, then these grades can be viewed as a population. We would expect that basic descriptive statistics associated with this population of grades, such as grade frequencies, mean, and standard deviation, to be relatively stable across time (see Hays, 1988, for a discussion of descriptive statistics). If we can also assume that a single evaluator has graded a relatively large and representative sample of the pilots in a particular fleet, not uncommon for large carriers, then an individual evaluator's grade distribution should resemble the statistical characteristics of the corresponding population of evaluators. Given a sufficiently large sample of observations (see Cohen, 1988, for details on sample size), deviations from the population distribution indicate that the evaluator is judging pilot performance differently from his peers.

There are two types of evaluator errors that can be identified from grade distributions of aircrew performance. First, an evaluator can grade either too strictly or too leniently. This type of error would be revealed when an evaluator's mean grade deviated significantly from the population mean grade. Second, an evaluator may fail to use the full range of the grade scale. Occasionally an evaluator will assign the grade "standard" for virtually all of his observations. This error can be readily identified when the evaluator's standard deviation of grades is significantly below that of the population standard deviation. Both types of error, abnormal mean and standard deviation, are examples of failing to appropriately apply the grade scale. Fortunately, grade scale use is under the direct control of the evaluator, and so simply bringing these discrepancies to the evaluator's attention should lead to improvement.

The mean and standard deviation do not provide a complete picture of an evaluator's grade distribution. Occasionally, the total frequency distributions of the evaluator and population should be compared. Because the number of distinct grade scale values is typically small (e.g., ≤ 5) for aircrew evaluations, it is easy to simply visually inspect distributional discrepancies. If needed, a measure of distributional congruency can be obtained by computing the mean absolute difference between frequencies (i.e., proportions) from an individual evaluator's distribution and a comparison distribution (Holt, Johnson, & Goldsmith, 1997; Williams, Holt, & Boehm-Davis, 1997). The comparison distribution may be derived empirically, such as the population distribution of grades for a fleet, or it may be specified on theoretical grounds, such as an expected proportion of grades.

A second major type of analysis of pilot performance data is a correlational analysis. The correlational structure of a set of performance items indicates which items vary together, presumably because the different items measure a common underlying skill. Nunnally and Bernstein (1994) describe the logic of the analysis; Martinussen and Torjussen (1997) give an example of the method applied to pilot selection. Our use of correlational structure to assess evaluators' performance is a variant of this application.

The logic behind the correlational analysis rests upon the fact that covariation in grades reflects observational skills of the evaluator. Two performance items will likely covary because they measure some common underlying piloting skill (e.g., use of flight automation). Only evaluators who recognize and discriminate among levels of aircrew performance will be able to generate the appropriate pattern of covariation. Some pairs of performance items may covary for

superficial reasons, such as they occur in the same context (e.g., phase of flight), and so even a less skilled evaluator might produce this type of covariation. But by examining the pattern of covariation over a large set of performance items taken from a large sample of observations, one could obtain an accurate prediction of an evaluator's grading pattern.

The first step in a structural analysis of grades is to compute a population correlation matrix. This matrix is the set of correlations between all pairs of items in the performance database computed across all pilots who have been graded on both items. Next, the same type of correlation matrix is derived for a single evaluator, where each correlation is based on grades given by that evaluator only. The similarity of the correlational matrices (population and individual) reflects the degree to which the individual's grading structure matches the population structure. This similarity can be quantified by computing the absolute difference between corresponding correlations, one from the population and the other from the individual evaluator.

A statistically significant difference between these two sets of correlations suggests that the evaluator is grading performance items differently from the norm[1]. Unlike grading errors uncovered with the mean or standard deviation, the meaning of a deviant correlational structure cannot be easily communicated to an evaluator, nor is the evaluator likely to have direct control over its production. In this case, the evaluator should receive additional training with a calibration session.

In summary, there are several types of statistics that can be computed from aircrew performance data that reflect, at least to some extent, how well an individual evaluator is grading. An obvious advantage of this type of feedback is that it is based on actual grading of pilot performance. A limitation is the requirement of a sufficiently large and representative sample of data to insure that statistical indices are reliable and meaningful. This condition may not be met in small carriers. However, whenever possible, carriers should monitor the performance database and provide routine feedback to evaluators regarding their grading practices. Over time this will improve and maintain the quality of aircrew evaluations.

Evaluator Training and Calibration Sessions

Calibration sessions provide the second source of data from which we can assess evaluators. Calibration sessions occur when multiple evaluators view and grade the identical aircrew performance. Grading exactly the same performance is necessary to assess the reliability of evaluator judgments.

At a most basic level, reliability reflects the consistency of a measurement (Anastasi, 1988). A single evaluator giving the same grade to the very same level of performance across multiple observations shows intra-rater reliability, and multiple evaluators giving the same grade to the same level of performance shows inter-rater reliability. Both are essential elements of quality evaluations.

Consistency by itself does not ensure that evaluations are sensitive or accurate. Sensitivity refers specifically to the degree to which observations covary with true changes in the attribute being measured. Accuracy reflects how well observations match established standards of assigning attributes to grade scale values. Both sensitivity and accuracy are components of reliability. We illustrate them with the following example.

Assume that three evaluators rated the same pilot's performance on several different performance items using a 5-point grading scale. Assume further that we knew the true level of performance for each of these items, which we refer to as the referent grade. Figure 1 shows that Evaluator A's grades generally covary with the true performance values (high sensitivity), whereas Evaluator B's grades deviate seemingly randomly from the true values (low sensitivity). Hence, Evaluator A is more sensitive than Evaluator B to variations in performance across the items, and so Evaluator A is more reliable in this sense.

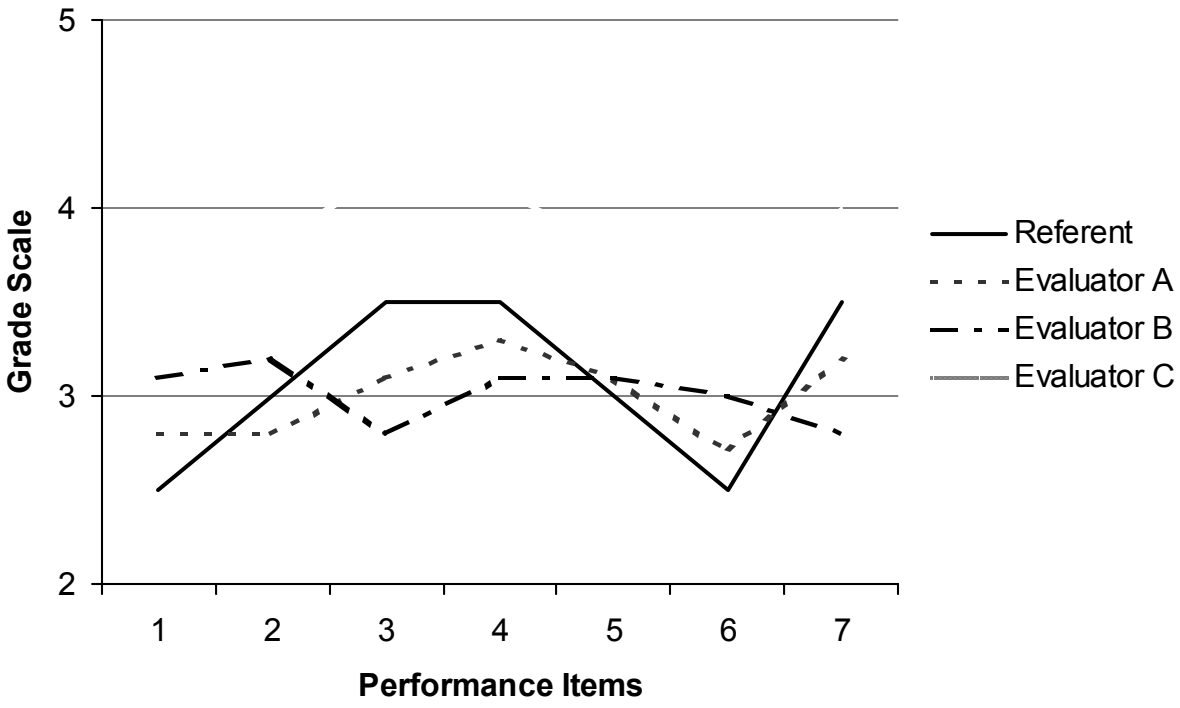

Figure 1. Examples of evaluator grading profiles showing different levels of sensitivity and accuracy.

However, sensitivity itself gives an incomplete picture of reliability as is illustrated in Figure 1 by Evaluator C. Evaluator C's grades covary almost perfectly with the true performance levels, but they lack accuracy in the sense that they consistently overestimate the pilot's true performance level. Evaluators must not only be sensitive, but their grades should systematically reflect established grading scale criteria by neither overestimating nor underestimating the true performance values. In sum, reliable observations must be both sensitive and accurate.

Quantifying Sensitivity: Correlational Measures. Because sensitivity depends on covariation of observed performance, it is measured with a correlational statistic such as the Pearson product-moment correlation. Evaluator reliability can be measured by correlating each evaluator's observations with every other evaluator's observations over some set of performance items and then averaging the resulting correlations to obtain a single evaluator reliability coefficient. This type of measure is referred to as inter-rater reliability (IRR). A single correlation can be obtained by averaging across the group of evaluators' correlations to reflect an

overall reliability coefficient (however, see Footnote 1). Thus, IRR reflects the degree to which a group of evaluators agrees with one another.

A second type of sensitivity measure can be obtained if the true performance grades are known for each performance item to be graded. Here each evaluator's grades are correlated with the true or referent grades to obtain a referent-rater reliability (RRR) coefficient. In practice, referent grades are based on the qualification standards associated with a particular task (e.g., ± 10 degrees heading deviation) to the observed performance. Later we describe how referent grades can be obtained. A group RRR coefficient can be computed by averaging the individual RRR scores. RRR reflects how closely each evaluator's ratings agree with some standard of grading. We believe that both IRR and RRR coefficients are useful indices of evaluators' ability to grade.

First, IRR gives evaluators feedback regarding how their grades compare with their peers. Such feedback can have immediate beneficial effects on an evaluator's grading performance by causing a deviant evaluator to conform to the group. However, knowing that evaluators agree with one another does not, by itself, guarantee that they are grading according to established standards of performance. In contrast, a high RRR score unequivocally implies that an evaluator's ratings are reflecting real changes in performance. It is in this regard that RRR may be considered a type of validity measure. The evaluator's ratings are valid in the sense that the variation in ratings reflects changes in performance that an airline considers relevant as defined by its qualification standards.

RRR measures sensitivity because it reflects the degree to which evaluators' ratings covary with true performance levels. This is not necessarily true of IRR. A high RRR coefficient implies a high IRR (i.e., if evaluators agree with the referent, they necessarily will agree with one another), but a high IRR does not imply a high RRR (i.e., evaluators may agree with one another but disagree with the referent grades). In practice, however, we have found RRR and IRR to be highly correlated, implying that the evaluators' grades generally agree with the referent grades. We have also found that the mean RRR value for a group of well-trained evaluators is typically in the vicinity of $\underline{r} = 0.80$ and is consistently higher than the mean IRR value. We have observed this pattern of findings for both technical and CRM performance items.

The fact that, on average, each evaluator agrees better with the referent grades than with other evaluators' grades corroborates the validity of the referent grades. Obtaining higher RRR

than IRR values likely occurs because, in any sample of evaluators, there are typically a few outliers whose ratings correlate poorly with the majority of evaluators. These outliers lower the mean correlation between all pairs of evaluators. Conversely, if IRR were found to be higher than RRR, this would bring into question the validity of the referent grades. In this case, one would want to analyze the performance items to determine if there were a subset of the items over which most of the disagreement occurred.

Finally, an additional advantage of RRR over IRR is that, because it is based on qualification standards, it provides an explicit training objective for evaluators. With appropriate training on the qualification standards and the grading scale criteria, evaluator's judgments should begin to show higher RRR values as their grades begin to conform to the referent grades.

Quantifying Accuracy: Mean Absolute Difference. As noted above, a high RRR does not necessarily imply accurate use of the grade scale. To supplement RRR, we devised a simple and direct way of measuring grade scale accuracy. We use a mean absolute deviation (MAD) coefficient to compute the average absolute deviations between grades given over a common set of performance items. MAD values range from a minimum of 0 (no deviations in the ratings) to a maximum value that equals the difference between the highest and lowest scale values (e.g., a 3 on a 1 to 4 grade scale). A standardized MAD (SMAD) score can be computed by dividing MAD by the maximum possible deviation and then subtracting this value from 1. SMAD ranges from 0 to 1, with better agreement corresponding to higher scores. Thus, SMAD allows meaningful comparisons to be made across different grade scales, and further it is scaled in the same direction as a correlation coefficient (i.e., higher scores mean better agreement). Although in practice we use SMAD, in this paper we refer to MAD.

To assess grade scale accuracy, MAD is computed between an evaluator's grades across some set of performance items and the corresponding referent grades. In studies we have conducted at several major airlines, we find that mean MAD values for experienced evaluators is typically quite small, generally deviating from referent scores by less than one quarter of a grade scale point. We also find that accuracy and sensitivity tend to covary; evaluators who are more accurate are also more sensitive. However, on occasion we do find some evaluators with good RRR and IRR values, but poor MAD values and to a lesser extent vice versa. The advantage of having multiple indices of grading performance is that evaluators can receive feedback specific to the nature of their problem.

In one sense MAD is a more fundamental measure than RRR or IRR because very good MAD scores imply not only good grade scale use, but good sensitivity too. In the extreme case, if MAD were perfect, both RRR and IRR would necessarily be perfect too. Further, as MAD scores become worse we might expect RRR and IRR to also approach a minimum level, but this is not necessarily so. As was illustrated in Figure 1, it is possible for a correlational measure such as RRR to be perfect while MAD is arbitrarily poor. For this reason it is necessary to compute both MAD and RRR to obtain a complete grading profile.

Another advantage of MAD over correlational measures is that it can be used to assess performance on individual items. As discussed so far, MAD is computed across items as RRR and IRR are, however it is also possible to compute deviations for an individual item across evaluators. Computing a MAD value for each performance item and then rank ordering items by this value shows clearly which performance items are difficult to grade. In group calibration sessions, we have found that this type of feedback generates much discussion about grading strategies and reasons for grading particular performance item.

The statistical reliability of a MAD value computed on items depends on the number of evaluators used in the computation. In contrast, correlations are computed across items for either a pair of evaluators or an evaluator and the referent grades, and so the statistical reliability of RRR and IRR is dependent on the number of items rather than on the number of evaluators. Consequently, MAD can be used in situations where a correlational analysis might not work, such as item analysis.

To summarize, we recommend several statistical indices be used to assess the quality of evaluator judgments, some based on distributional characteristics of pilot grades, and others used in the context of calibration sessions. We introduced two statistics, RRR and MAD, that assess important, yet independent, aspects of rater reliability, sensitivity and accuracy, respectively. These statistics along with the more traditional inter-rater reliability should provide a relatively complete grading profile. Next we discuss the design and implementation of training and calibration sessions.

Evaluator Training and Calibration Sessions

Here we draw upon our experience in doing training and calibration sessions over the last few years at several major carriers in the United States. Most of our findings and

recommendations come from informal and naturalistic observations and so should be viewed tentatively. Nevertheless, others engaged in evaluator training may find them useful.

Training and calibration sessions routinely occur when groups of instructors and evaluators (usually from a common fleet) assemble for standards meetings where issues pertinent to evaluation are discussed. By training and calibration session, we mean specifically the observing and grading of aircrew performance and the subsequent feedback phase that is associated with this task.

The Role of the Evaluator

We distinguish between the evaluative versus diagnostic role of assessment. Evaluators naturally focus more on the evaluative aspect of assessment. This tendency is understandable given that their first priority is to ensure pilot proficiency. But quality training hinges upon accurate diagnosis of pilots strengths and weaknesses. It is through careful evaluation of pilot performance that carriers know which training components are working and which are not. Evaluators who acknowledge their role in the training mission are likely to support the implementation of assessment tools and methods.

Artificial Nature of the Grading Sessions

A potential limitation of calibration sessions is their artificial nature. Judgments of aircrew performance are solicited out of their normal context and without the full complement of information normally available to the evaluator in the simulator. On the other hand, these contextual factors may bias evaluator judgments and from this perspective calibration performance provides a better estimate of an evaluator's true grading skills.

Line Operational Evaluations

In an LOE, an evaluator is asked to judge performance at both the event set level and at a more detailed performance level called observable behaviors. An event set is defined as a meaningful segment of flight, and may correspond to a particular phase of flight or to a special occurrence within the flight, such as an anomaly. Evaluators usually grade both technical and CRM skills at the event set level.

In addition, there are usually multiple observable behaviors graded within an event set. The observable behaviors reflect both technical and CRM tasks. These items vary greatly across carriers from relatively general categories of performance (e.g., effectively communicates) to

highly specific behaviors (e.g., de-icing procedures). LOE grade sheets may include 30 or more observable behaviors in addition to the event set items to be graded. Not surprisingly, evaluators are challenged to perform this extensive assessment while at the same time operating the simulator. One of the most direct ways to assist evaluators in grading LOEs is to design a good LOE grade sheet.

Designing Grade Sheets

The design of a grade sheet plays a critical role in evaluators' ability to grade aircrew performance, particularly for the LOE. Unfortunately, good grade sheets are difficult to design. Part of this difficulty stems from the fact that there are many ways to fail.

A central question about LOE grade sheets is what types of performance items (observable behaviors) to include. LOEs are intended to assess both CRM and technical skills, and so separate performance items for each category are included. If observable behaviors are intended to diagnose specific deficiencies in the training program, then they must be sufficiently detailed. Further, they need to be linked to the tasks and subtasks defined in the training curriculum. Unfortunately, we know of no empirical data to guide carriers in writing these items. At what level of detail can evaluators reliably grade performance? How many items are needed to reliably assess performance on a particular skill category? What particular subskills (e.g., decision making, situational awareness, workload management, etc.) underlie CRM and what types of performance items allow these subskills to be reliably measured?

Further, how important is the wording of observable behaviors? In one study at a major carrier, we found that relatively simple changes in the phrasing of performance items led to substantial improvement in interrater reliability. As an example, we found high disagreement among evaluators grading the same aircrew on "Engine-out missed approach procedures." After rewording the item to "Performs engine-out precision approach procedures and maneuvers IAW POM," evaluator agreement increased 63% on a subsequent session.

Instead of using specific observable behaviors to diagnose deficiencies, some carriers rely more on reason codes. A list of reasons justifying the assignment of a grade (e.g., poor workload management, inadequate knowledge of some subsystem of the aircraft, etc.) is provided on the gradesheet. The evaluator is asked to indicate a reason for grades that are exceptionally high or low. Do reason codes allow as fine of a diagnosis of proficiency as observable behaviors? Is

evaluator reliability as good with reason codes as with ratings on performance items? These and related questions need to be answered in future work.

Grade scales range from two-points (e.g., pass/fail) to five points. Is there an optimal number of grade scale levels? There is little value in using a 10-point scale if evaluators are capable of reliably discriminating only five levels of performance. Conversely, if evaluators are capable of discriminating five levels of performance, valuable discriminative information is lost with a 2-point scale.

We suggest using the same grade scale for all performance items. Some airlines use a 3-point scale used for grading critical maneuvers and then use a 4- or 5-point scale used for grading items in an LOE. Even within an LOE, different scales are sometimes used to grade different types of items. Employing a consistent grade scale across types of items facilitates the evaluator's task and further allows easier interpretation of performance data.

How important are grade-scale labels? The distribution of grades assigned by evaluators is affected by the choice of labels. For example, with a 5-point scale (where a 5 indicates a high level of performance) a grade of 3 could either be labeled as “standard level of performance” or it could be labeled as “below standard level of performance.” When a grade of 3 is below standard, the vast majority of grades are 4’s. Consequently, there is little discrimination in levels of passing (i.e., grades of 4 or 5, with very few 5’s). Carriers must decide if it is more important to discriminate levels of below-standard performance or levels of above-standard performance.

In reality, the number of scale points and scale labels are often determined by corporate history and may be resistant to change even in the face of empirical data. At the very least, carriers should monitor their grade distributions to determine how well their grading practices are actually reflecting the desired levels of discrimination in performance.

Finally, as a general rule, grade sheets need to be designed with the user in mind. Ideally, evaluators themselves would help design the grade sheet. Unfortunately, evaluators may not have the same purpose in mind when making decisions about properties of a grade sheet as say a training coordinator. As mentioned above, evaluators are generally more concerned with the evaluative function than the diagnostic function of a grade sheet. One strategy that we have found to be effective is to solicit evaluator input, but do so by offering choices between alternative methods of implementing aspects of the grade sheet.

Establishing Referent Grades

Earlier we discussed the use of referent grades in assessing evaluator performance. An important part of designing a calibration session is the assignment of referent grades to each performance item to be graded. This should be done by a group (i.e., at least 4 or 5) of supervisory level evaluators. Ideally, supervisors would first assign grades independently and then later resolve any disagreements as a group. Differences that are not easily resolved may suggest that (a) the relevant behavior is not clearly represented in the video, (b) the performance item is not clearly stated on the grade sheet, or (c) the link between the item and the appropriate qualification standard is not clearly defined. If disagreements persist, the item should be removed from the grade sheet.

Airlines have developed explicit qualification standards for most technical tasks and subtasks. These qualification standards serve as the basis for pilot training and also guide the evaluation of aircrew performance. In addition, there are usually clearly stated grade scale criteria that describe how degrees of deviation from the qualification standards are to be mapped onto particular grades (e.g., a heading deviation that goes beyond a ± 10 degrees limit but is quickly corrected results in a 3 on a 4-point scale). If evaluators are to be trained and assessed on grading aircrew performance, including CRM skills, well-specified qualification standards and grade scale criteria are needed. Unfortunately, these are rare or nonexistent for CRM tasks. Nonetheless, we find that experienced evaluators are able to agree upon referent grades for CRM performance items.

Creating a Performance Video

Two important criteria for creating videos of aircrew performance for use in calibration sessions are that the videos must be realistic and they must contain different levels of performance. One means of achieving realism is to use videos of actual crews performing an LOE or flying a critical maneuver. The use of these videos would likely require approval from the crew and the pilot union. Selected sections of the video would be used for the calibration session. Consequently, they may not reflect performance of the crew on the LOE. Finally, the video quality of these tapes is often poor.

Alternatively, we have found that it is possible for pilots to enact a scripted performance scenario. An experienced crew is instructed to achieve a certain level of performance during part

of an LOE or a maneuver. High quality videotapes can be produced this way with careful planning, filming, and editing of the video. In either case, the flights must appear realistic to the evaluators. If deemed critical, evaluator judgments of video realism could be collected as part of a calibration session to validate the video's realism.

In a calibration session, evaluators normally grade several video clips. Scenarios with different levels of aircrew performance are needed to assess evaluators' sensitivity to changes in the quality of aircrew performance. For example, if performance were graded on a 4-point scale, then scenarios representing at least three levels of the grade scale should be used. It is particularly difficult to create scenarios depicting weak or failing performance such that the deficiencies are not too obvious.

Finally, it is important that the actual aircrew performances to be evaluated are clearly represented on the video. Conversely, if the video has been scripted to omit certain behaviors, it should be clear from the context of the video where the behavior would have occurred. Once the video and grade sheet are developed, a group of expert evaluators should observe and grade the scenarios to determine that the performances did occur in the flight scenario. This step can occur during the collection of referent grades.

Evaluator Feedback

Most of the evaluator training occurs during the feedback phase of a calibration session. In the case of a group session, each evaluator receives a report showing the evaluator's individual ratings along with a summarization of the group's ratings. Summary measures of grading performance are provided as feedback including RRR, IRR, and MAD. An evaluator supervisor then discusses and interprets the statistics for the group.

An important type of feedback is to simply show the list of the performance items that resulted in the average highest and lowest overall agreement with the referent grades. The high agreement items illustrate concretely that it is possible for the evaluators to uniformly agree with the referent and with each another. In discussing the lowest agreement items it is useful to provide the qualification standards that were the basis for generating the referent grades and to review the grading scale criteria.

It is also possible to carry out calibration sessions for individual evaluators. Here a single evaluator views and grades digitized videos on a personal computer and then receives feedback

similar to a group session[2]. In addition to the flexibility of when and where calibration occurs, individual calibration allows a supervisor to select particular types of performance evaluations (e.g., critical maneuvers vs. LOEs) for an evaluator to grade. Further, training feedback can be more specific by allowing the evaluator to examine qualification standards for only those performance items he graded incorrectly, or to replay critical parts of the flight that he graded incorrectly.

An even more elaborate type of evaluator feedback that can be given during a calibration session is based on frame-of-reference training (Baker & Mulqueen, 1999; Baker, Mulqueen, & Dismukes, 1999; Baker, Mulqueen, & Dismukes, in press; Bernardin & Buckley, 1981). In this case a gold standard, which explains in detail the rational for the assignment of the referent grade, is developed for each performance item. Gold standards are not simply more detailed qualification standards, but they go much further by discussing the role of the specific flight context in the assignment of a grade to a performance item. A current study at a major airline is examining the effectiveness of gold standards in improving evaluator performance.

Summary and Future Directions

The evaluation of human performance in highly skilled jobs, such as commercial flying, is still largely a human activity, and as such, retains an inherently subjective component. These evaluations play critical roles in deciding the careers of pilots, ensuring the safety of commercial flying, and in designing and modifying training programs. Hence, it is critical that these performance judgments be of the highest quality. Fortunately, much can be done to ensure that these human judgments are accurate and reliable.

Psychometrics

We have suggested that there are two basic components of quality evaluations: sensitivity to changes in performance levels and accuracy in the use of the grade scale. We defined two statistical measures, RRR and MAD, that quantify these components. Both of these measures compare an evaluator's performance ratings to a set of referent ratings that reflect the qualification standards under which pilots are trained and against which evaluators judge. Deviations from these referent grades have clear implications for training and calibrating evaluators.

A topic for future research is to consider statistical measures of evaluator performance based on ordinal methods. There are several characteristics of rating data that suggest ordinal

measures should be used. First, the evaluator ratings themselves may not contain more than ordinal-level information. In other words, the difference in actual performance between grades of 2 and 3 is likely to be different from the difference in performance between grades of 3 and 4. Yet, standard statistics of the sort commonly used to analyze grade data assume interval-level information. Second, grade distributions are never normally distributed and instead are often highly skewed and/or contain little variance. Third, we sometimes wish to compute a statistic on a small number of items, where variances and covariances are not very stable. Ordinal measures have several advantages over variance-covariance based measures under these conditions (Cliff, 1996).

Training and Calibration Sessions

There are several questions that should be addressed in future studies aimed at an empirical validation of evaluator calibration sessions. First, assuming that evaluators do improve their judgments during calibration sessions, how long does this improvement last? Once calibrated, does judgment accuracy decay over time? Second, how well do calibration sessions based on specific flight segments and maneuvers generalize to other flight situations? When evaluators are calibrated under specific events and phases of flight, will this learning transfer to different phases of an LOE? Third, what effect do evaluator demographics (e.g., months as an evaluator, hours of flying, years with carrier, age, etc.) have on evaluator performance, and what effect do these characteristics have on the effectiveness of evaluator calibration sessions? Fourth, what types of grading feedback are most effective in training evaluators? Is the more extensive effort required to develop gold standards justified? Finally, and most importantly, do the positive effects of calibration sessions result in improvements to actual aircrew evaluations? Can we see a pre-post calibration improvement in grading as reflected in actual aircrew performance data?

Finally, we suspect that for calibration training to remain effective it will need to be extended beyond occasional (e.g., annual) group calibration sessions. An obvious means of accomplishing this would be to allow evaluators to perform individual calibration sessions. With the implementation of a software calibration tool at a couple of major carriers, we plan to study how this type of additional training affects both group calibration performance and actual evaluation of aircrew performance.

References

Anastasi, A. (1988). Psychological testing; New York: Macmillan Publishing Co.

Arvey, R. D. & Murphy, K. R. (1998). Performance evaluation in work settings. Annual Review of Psychology, 49, 141-68.

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

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

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

Baker, D. P. & Salas, E. (1996). Analyzing team performance: In the eye of the beholder? Military Psychology, 8, 235-245.

Bernardin, H. J., & Buckley, M. R. (1981). A consideration of strategies in rater training. Academy of Management Review, 6, 205-212.

Burke, E., Hobson, C., & Linksy, C. (1997). Large sample validation of three general predictors of pilot training success. International Journal of Aviation Psychology, 7, 225-234.

Cliff, N. (1996). Ordinal methods for behavioral data analysis. Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2^{nd} ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Damos, D. L. (1996). Pilot selection batteries: Shortcomings and perspectives. International Journal of Aviation Psychology, 6, 199-209.

Dwyer, D. J., Oser, R. L., Salas, E., Fowlkes, J. E. (1999). Performance measurement in distributed environments: Initial results and implications for training. Military Psychology, 11, 189-2125.

Federal Aviation Administration (1991). Advanced Qualification Program. FAA Advisory Circular 120-54.

Fisher, R. A. (1925/1990). Statistical methods, experimental design, and statistical inference. Oxford: Oxford University Press.

Hays, W. L. (1988). Statistics. (4th ed.) NY: Holt, Rinehart and Winston.

Hoermann, H. & Maschke, P. (1996). On the relation between personality and job performance of airline pilots. International Journal of Aviation Psychology, 6, 171-178.

Holt, R. W., Johnson, P. J., & Goldsmith, T. E. (1997, June). Application of psychometrics to the calibration of air carrier evaluators. Paper presented at the meeting of the Human Factors and Ergonomics Society, Albuquerque, NM

Hubbard, D. C, Rockway, M. R., Waag, W. L. (1989). Aircrew performance assessment. In Jensen, R. S. (Ed.) Aviation Psychology (pp. 342-377). Brookfield, VT: Gower Publishing Co.

Jensen, R. S. (1989). Aviation psychology. Brookfield, VT : Gower Publishing Co.

Martinussen, M. & Torjussen, R. (1997). Pilot selection in the Norwegian air force: A validation and meta-analysis of the test battery. International Journal of Aviation Psychology, 8, 33-45.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory. New York: McGraw-Hill, Inc.

Salas, E., Prince, C., Baker, D. P. & Shrestha, L. (1995). Situation awareness in team performance: Implication for measurement and training. Human Factors, 37, 123-136.

Williams, D. M., Holt, R. W., Boehm-Davis, D. A. (1997). Training for inter-rater reliability: Baselines and benchmarks. In the proceedings of the Ninth International Symposium on Aviation Psychology, Ohio State University, (Richard S. Jensen and Lori A. Rakovan, Editors) Volume 1, pp 514-520

Footnotes

[1] It is recommended that correlation coefficients first be transformed to Fisher Z scores before combined or tested for statistical significance (Fisher, 1990).

[2]A software tool is available to carry out individual calibration sessions. In addition, the tool allows data from group sessions to be entered, performs the requisite statistical analyses, and prints feedback reports. Contact the first author for further information regarding the software.

Inferential Statistics and Sample Size: Applications to Air-carrier Training and Assessment

Introduction

Airlines along with other industries are increasingly relying on statistical information to guide important corporate decisions. In the area of training and assessment, airlines need to answer such questions as whether the current training curriculum is adequate, and whether pilots are maintaining a proper level of proficiency between periods of retraining. Wrong answers to these types of questions can have serious and adverse consequences. Fortunately, by collecting quality training and assessment data and by applying modern statistical methods to guide decisions, we can minimize these errors.

In answering questions about the quality of training and pilot performance we generally wish to infer something about a population of individuals based only on a sample from the population. For example, we might want to know how well the pilots from the B757 fleet perform on engine failure after V1. Sampling the entire population of pilots from this fleet, say for example 300, is simply not practical. Instead, we select and test only a sample of these pilots. We then infer the performance of the entire fleet based on the performance of the observed sample. The branch of statistics that allows us to infer characteristics of a population from sample observations is called inferential statistics.

A central issue within inferential statistics is deciding how large of a sample is needed in order to reliably estimate values in a population from the sample. This issue is the main concern of this section of the chapter.

Hypothesis Testing Framework

Before we examine specific methods for choosing appropriate sample sizes we first need to review the general idea of hypothesis testing. Hypothesis testing is the logical framework for making statistical decisions. The basic question we wish to answer in inferential statistics is whether some observed sample value is different from some known population value. For the sake of argument, we assume that it is not different and call this assumption the null (i.e., no difference) hypothesis, which we designate by the symbol Ho. The alternative hypothesis, Ha, states that the sample value truly is different from the population value. Under the hypothesis-testing framework, these are the only two possibilities: either the null hypothesis is correct or the alternative hypothesis is correct. We never know with certainty which hypothesis is the true

state of the world. But based on sample data we choose one of these two hypotheses. Inferential statistics helps us make the optimal decision.

Our decision will always be either to reject or to accept the null hypothesis, and once we have decided what to do, we will be either right or wrong. There are two ways of being right and two ways of being wrong. Table 1 illustrates these four possibilities. If we reject Ho when we should have accepted it, we make what is called a Type I error (upper left cell). The probability of a Type I error is equal to α. If we retain Ho when we should have rejected it, we make a Type II error (lower right cell), and it has a probability of β.

In a standard research setting, special privilege is given to the null hypothesis because it represents the status quo. Science is conservative by nature, and so tries to avoid claiming that there is something new when in fact there is not. For this reason Type I errors are especially avoided. To guard against these false positives, we set α to some acceptably low limit, usually around .05. This means that there is only a one in 20 chance of incorrectly rejecting Ho. Committing a Type II error is not viewed as grievous as a Type I error. The consequences of failing to believe that some new thing is significantly different from the old when it is not (false negative) is assumed to be less serious than believing that something is significantly different when in fact it is not (false positive). As we will see shortly, this type of bias may not hold in the applied world of training and assessment.

There are also two ways of being correct. From a research perspective, the most favored of these two outcomes is to correctly reject Ho (upper right cell). Here we claim that we have discovered something new, and indeed we have. Associated with this outcome is the power of the test and its probability is $1-\beta$. We can also be correct by retaining Ho (lower left cell). Although correct, this outcome is less desirable because nothing new has been found.

Table 1. Possible outcomes of hypothesis testing framework.

		True State of World	
		Ho is true	**Ho is false**
Decision	**Reject Ho**	Type I error (α) (false positive)	Power ($1-\beta$)
	Retain Ho	Correct decision ($1-\alpha$)	Type II error (β) (false negative)

In applied settings such as training and assessment, the relative advantages and disadvantages of these four possibilities may be different from traditional scientific research. To consider a concrete example, assume that we know that the population performance level for a particular fleet and for a particular maneuver was a mean grade of 3.0 on a 1 to 4 grading scale, where a "1" is poor and a "4" is good. Assume further that we obtain a sample of pilots from this fleet and observe that their mean performance on the maneuver is 2.7. Do we now conclude that the performance of the whole fleet has degraded from 3.0? Or could this lower mean have occurred simply by chance because we examined only a portion of the pilots?

In this example the null hypothesis corresponds to the belief that the performance of the fleet has not changed (i.e., status quo); the alternative hypothesis says that it has. Unlike scientific research, in this case there does not seem to be a compelling reason to favor Ho over Ha. It seems just as important to recognize that there has been a change in fleet performance (perhaps even more so) than to recognize that performance has not changed.

Consider again the two types of errors we can make. If based on the performance of the sample we conclude that the performance of the fleet has changed when it has not, then we have made a Type I error. Here we would probably require additional training when it was not needed. This would certainly be an error and cost the company unnecessary resources. On the other hand, if we retain Ho when Ha is true, then we have made a Type II error. In this case, the fleet needed additional training on the maneuver being examined but we failed to recognize it. The consequences of this error could be more serious than the first type. Clearly, it is just as important to guard against false negatives (maybe more so) as false positives in this setting. Thus, when using the hypothesis-testing framework with training and assessment data we will examine both false positive and false negative error rates. As it turns out, the particular error rates that we accept have a direct bearing on the size of the sample we need to obtain in order to assess the population.

Sample Size and Sampling Error

So far we have described the logic of the hypothesis-testing framework and saw how it could be applied to making decisions about training and assessment data. We turn next to the issue of sample size and a closely related concept, sampling error. First, sampling error is simply the inaccuracies that arise when we measure individuals from a sample rather than

measuring the whole population. To continue with our previous example, if we sample 50 of the 300 B757 pilots and test them on an engine failure after V1 we might obtain a mean score of 2.7. If we could have tested all 300 pilots we would have found that their mean performance was 3.0. Sampling error is the difference between the scores we observe in a sample and the scores that truly exist in the population.

There are two major factors that influence sampling error. The first is how representative our sample is of the population. If the sample's characteristics match well those of the population, then we have a representative sample, and sampling error is minimized. The best way to ensure that we obtain a representative sample is by random sampling. Random sampling occurs when every individual in the population has an equal change of being selected for the sample. Unfortunately, in real-world applications such as pilot evaluation it is virtually impossible to sample randomly because of the constraints on scheduling, time since previous evaluation, etc. The second major influence on sampling error is sample size; the larger the sample the smaller the sampling error. Notice that if our sample size were actually the whole population, then by definition, there would be no sampling error at all. Our "sample" would perfectly mirror the population. As sample size decreases, there is an increasing chance for the pilots in the sample to differ from the population, and so for sampling error to increase. The question then becomes, How big of a sample do we need in order to be confident that our findings accurately reflect the population?

Choosing a Sample Size

Choosing an appropriate sample size depends on (a) which statistical test we use; (b) the population variance of the statistic being measured; (c) the effect size in the population we wish to be able to detect; (d) an acceptable level of α; and (e) an acceptable level of β. We will examine each of these in turn. The appropriate statistical test will be determined by which statistic we are measuring (e.g., mean performance) and what hypothesis we are testing. With assessment data, our statistic is often either a mean grade (which we will designate by $\bar{X}$) or a proportion of pilots passing some evaluation. We will say more about statistics later, but for now assume that we are interested in mean performance. Also, our statistical test will either be a one-tailed test or a two-tailed test. In a one-tailed test, we assume that we know that the hypothesized population mean under the alternative hypothesis (μa) is greater or smaller than the hypothesized mean under the null hypothesis (μo). In a two-tailed test we recognize that it could

be either greater or smaller. Generally, we will want to perform a two-tailed test because we rarely know for certain what direction the outcome will be. However, two-tailed tests are more conservative than one-tailed tests. Conservative, in this case, means that we would need a larger sample for a two-tailed than a one-tailed test to maintain the same error rates.

The population variance, which we designate by σ, is a measure of how much variability exits in the scores of the pilots. Do the scores vary widely across the values of 1 to 4 (high variance) or are they generally centered on say "2"s and "3"s (low variance)? We usually estimate σ from sample data, although there may be times when we know this value for the fleet in question. As population variance increases, our sample size needs to increase. Notice that in the extreme (and unrealistic) case where there is no variance in the population scores, we would only need to sample one case in order to determine if the population had changed.

The population effect size is the difference in performance between the parameters assumed under Ho and Ha. When testing sample means, the effect size is simply the difference between μo and μa. The hypothesized value under Ho will likely come from prior knowledge of fleet performance. The hypothesized value under Ha will be based on how much of a difference from μo that we deem as unacceptable. For example, we might believe that one-quarter of a grade point is truly significant. If we believe that fleet performance has decreased one-quarter of a grade point then we will take some type of action (e.g., decrease the interval between retraining). Finally, we need to be willing to state acceptable levels of error, both for false positives and false negatives. Choosing acceptable population effect sizes and error rates will ultimately be a corporate decision, although there are guidelines available from the statistical and research community.

Again to make these ideas more concrete, assume that we are going to test a sample of B757 pilots on engine failure after V1. We know from previous testing that the fleet as a whole scored 3.0 on this maneuver (i.e., $\mu o = 3.0$) . We wish to be able to detect a change in performance of one-quarter of a grade (i.e., $\mu a = 2.75$ or $\mu a = 3.25$). We also know from previous data that $\sigma = 0.60$. Further, assume that we are willing to accept a false positive rate of $\alpha = 0.20$ and a false negative rate of $\beta = 0.10$. Our null and alternative hypotheses would be formally stated as follows: Ho: $\mu o = \mu a$ (fleet performance on the maneuver has not changed) and Ha: $\mu o \neq \mu a$ (fleet performance on the maneuver is different from the assumed value).

What is the minimum number of pilots we would need to sample from the fleet in order to determine, within the stated error rates, that fleet performance had changed?

Questions about required sample size are answered from either specialized computer programs that generate minimum sample sizes or from published tables. For example, Table 2 gives actual minimum sample sizes needed to detect a quarter point change in grades for different levels of error rates. To answer our question above, we see from Table 2 that we would need to sample 38 pilots. One problem with using tabled values to obtain minimum sample sizes is that published tables may not give the needed values for the specific set of parameters at hand. A computer program that dynamically generates minimum sample sizes for given sets of parameters is a better solution.

To continue with the example, assume that we have now sampled and tested 38 pilots from the B757 fleet on engine out after V1 and we find that the mean grade is significantly lower than 3.0. Based on the appropriate statistical test (e.g., a two-tailed z test) we decide to reject Ho and conclude that the fleet's population mean grade on engine out after V1 has truly deteriorated. Are we absolutely certain of this? No, in fact there is a one in five chance ($\alpha = 0.20$) that we have made a false positive decision. On the other hand, assume that the mean grade from the sample was not sufficiently extreme to reject Ho. There is now a one in ten ($\beta = 0.10$) chance that we have made a false negative decision and concluded that fleet performance has not degraded when in fact it has.

The point of the example is to illustrate that there is always uncertainty associated with statistical decisions. The primary means that we have to reduce uncertainty is to increase sample size. We see from Table 2 that if we had sampled 62 pilots instead of 38, then our false positive rate would have been $\alpha = 0.10$ and our false negative rate would have been $\beta = 0.05$. Obviously there are costs involved in obtaining larger samples. The decision about what are acceptable sample sizes and error rates must ultimately be made by a data analyst who is cognizant of all of the relevant considerations.

Table 2 Minimum sample sizes for detecting a difference of .25 grade points with a one-sample z-test and assuming $\sigma = 0.60$.

		β

		0.05	**0.10**	**0.20**	**0.30**
α	**0.05**	75	61	45	36
	0.10	62	49	36	27
	0.20	49	38	26	19
	0.30	41	31	20	14

Some Related Sampling Concerns

We now briefly discuss a few other issues that are important in sampling and inferential statistics and seem particularly relevant for airline applications. Our intent here is to simply alert the reader to these issues rather than providing any type of comprehensive coverage.

Aggregating Samples Across Time

Samples of pilots are necessarily obtained within some unit of time. We might test 30 B757 pilots on engine failure after V1 during a single month. Performance of this sample would be an estimate of how well the fleet performs during this month. We might test another 30 B757 pilots on the same maneuver on the following. Can we now combine these two samples and perform a single statistical test on the performance of the combined sample of 60 pilots? As discussed above, by increasing our sample size our estimate of fleet performance will be more accurate and our false positive and false negative error rates would be lower. However, the answer to question is the proverbial "it depends." If the performance of the population has remained stable, then we can reasonably combine the samples from the two months. But if fleet performance has changed over time (and detecting such changes is the very purpose of assessing in the first place), then we need to be cautious about combining samples. Ideally, we would sample within as small of a unit of time as possible. In reality, we must sample and test pilots within the constraints of the operational world. Combining samples of pilots across two or even more consecutive months may be necessary. But what about combining samples taken 12 months apart? It is obviously unreasonable to combine samples over a time period as long as the period within which we are attempting to test for changes. Once again, these types of questions will have to be answered by a data analyst who is knowledgeable about the whole set of relevant considerations.

Sample Size and Different Statistical Tests

In our examples we have used average grades as the statistic for assessing performance. The sample mean is undoubtedly the most commonly used statistic and is arguably the single best single summary index of group performance. However, there are times when we may want to know other characteristics of group performance. Another common statistic is the proportion of a group that passes an assessment. For example, we might know that in the past 95% of the B757 fleet passed a first-look evaluation of engine failure after V1. We now want to know if this pass rate has changed. To find out we would obtain a sample of pilots from the fleet, administer the engine failure after V1, and measure pass rate. We could then perform a statistical test on this proportion in much the same way as for a sample mean. The same basic ideas of hypothesis testing apply here as before including false positive and false negative error rates. However, the actual sample sizes needed to achieve the same levels of error rates for a test on sample proportions will likely be different from a test on sample means. In fact, minimum sample sizes need to perform a statistical test on sample proportions may be much larger than those for corresponding sample means.

Sampling from Relatively Small Populations

Most applications of inferential statistics involve sampling from very large, and for all practical purposes infinite, populations. Drawing samples without replacement from such populations has a negligible effect on the sampling distribution. However, occasionally we sample from relatively fixed and small populations, as is the case in airline training and assessment. In such cases, there exists a correction to the calculation of the sampling variance of the mean. The relevance of this correction to the present discussion is that by using the correction the sampling variance of the mean is reduced and hence the required sample size is also reduced accordingly. The formula for calculating the corrected sampling variance is:

$$\sigma_M^2 = \left(\frac{N - n}{N - 1} \right) \frac{\sigma^2}{n} \qquad \text{Equation 1}$$

where N is the size of the population, n is the size of the sample, and σ^2 is the variance of the population. To illustrate the effect of this correction, we show in Table 3 the minimum sample sizes needed to detect a difference of .25 grade points and for fixed values of $\sigma = 0.60$, $\alpha = 0.05$, $\beta = 0.05$ and assuming different population sizes. As the size of the population decreases, the

minimum sample size also decreases. Looking back at Table 2 we see that the minimum sample size for the same set of values without the correction is 75.

Table 3 Minimum sample sizes to detect a difference of .25 grade points, assuming $\sigma = 0.60$, $\alpha = 0.05$, $\beta = 0.05$, and using correction given in Equation 1.

	Population size			
	100	200	300	500
Sample size	43	55	60	65

Summary

In a training program that is truly proficiency based, such as AQP, the collection and analysis of quality data is of paramount importance. Decisions about the adequacy of training will ultimately be answered by empirically derived assessment data. Many of the questions that arise about our training program can be answered by simple descriptive statistics: How many pilots were trained this month? What was the average grade on engine failure after V1? and so forth. However, at some point questions will arise that require drawing inferences about some larger group of individuals (e.g., a single fleet) from observations made on only a sample of those individuals. Has fleet performance on a particular maneuver degraded? Can we reduce the retraining interval for the B757 fleet? These and related types of questions will necessarily involve inferential statistics.

In this section of the chapter we provide a brief discussion of inferential statistics and the types of issues that arise when we perform statistical tests. We especially focused on the question of selecting an appropriate sample size. Our discussion has been elementary; there are many more technical issues that could be addressed but are beyond the scope of this chapter. Fortunately, there are many good introductory and advanced statistics books that address these issues, some of which we have listed below in the bibliography.

Bibliography

Bradley, J. V. (1976). *Probability; decision; statistics*. Englewood Cliffs, NJ: Prentice-Hall.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev ed.). New York: Academic Press.

Hays, W. L. (1988). *Statistics* (4th ed). New York: Holt, Rinehart and Winston.

Pagano, R. R. (1994). *Understanding statistics in the behavioral sciences*. (4th ed.). Minneapolis: West.