The Importance of Quality Data in Evaluating Aircrew Performance

Peder J. Johnson & Timothy E. Goldsmith

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.

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.

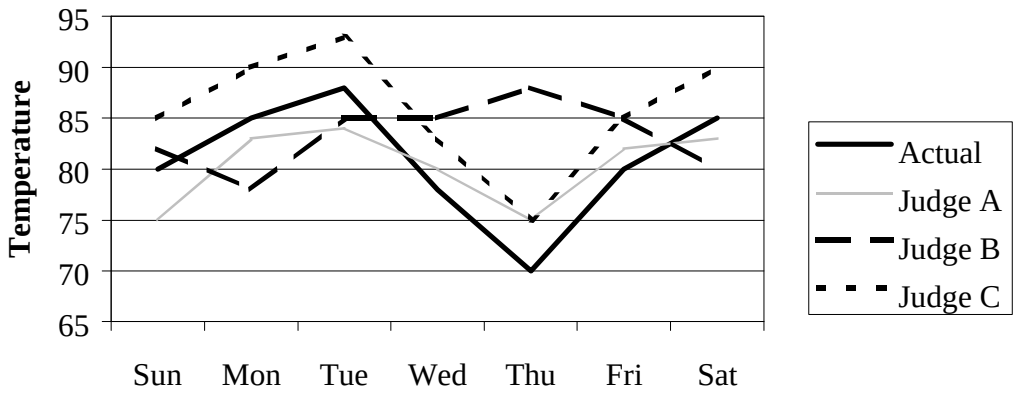

Figure 1. Graphs depicting different Judge's deviation from actual temperature

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.

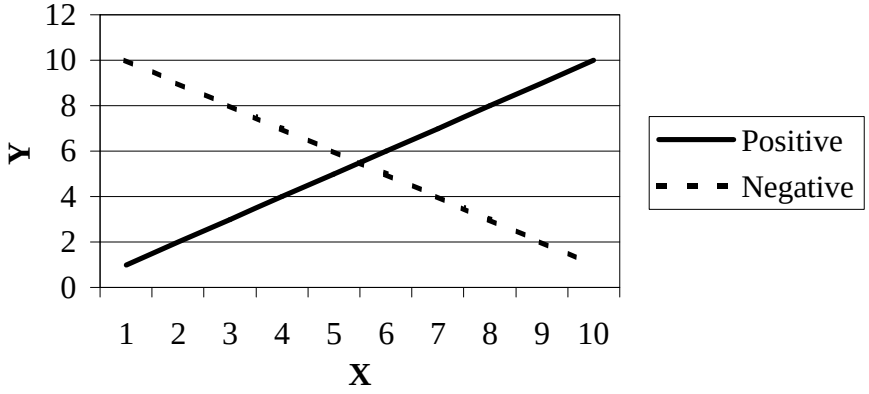

Figure 2. Illustration of positive and negative correlations

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,

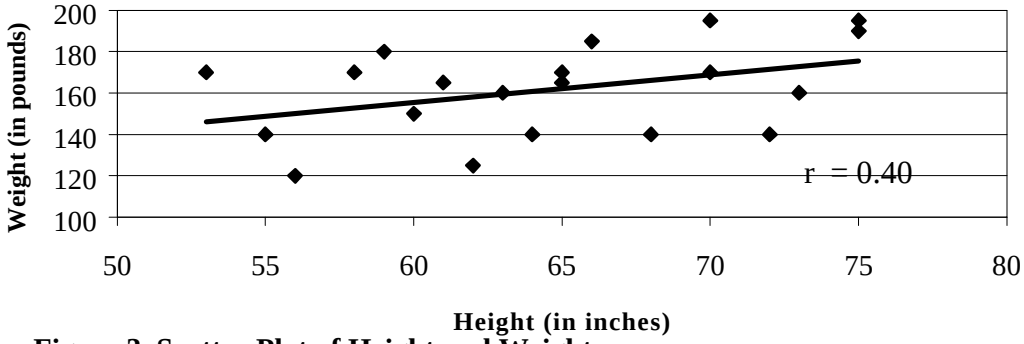

Figure 3. Scatter Plot of Height and Weight

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.

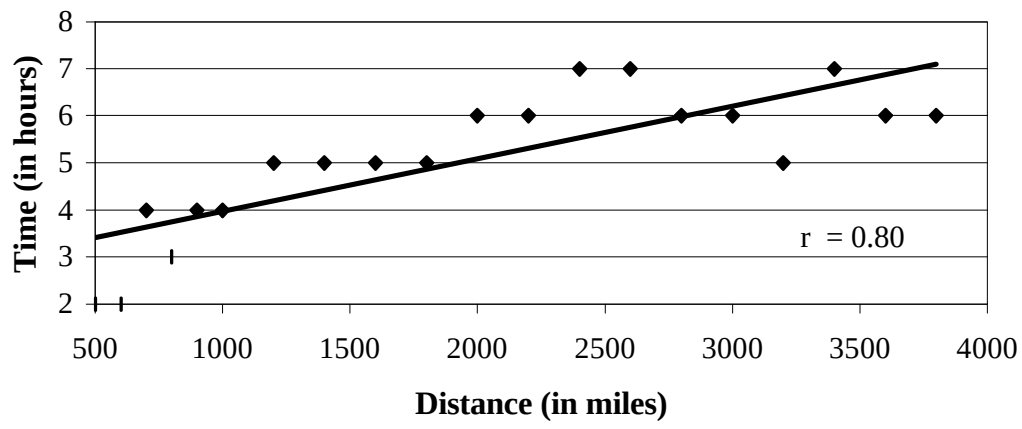

Figure 4. Scatter plot showing covariation between time and distance

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 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.

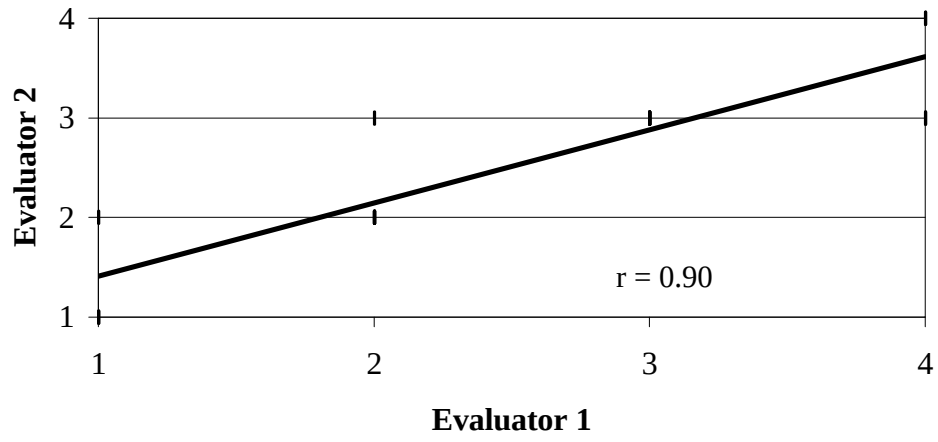

Figure 5. Two Evlauators' ratings of 20 performance indicators on a 4-point scale.

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.

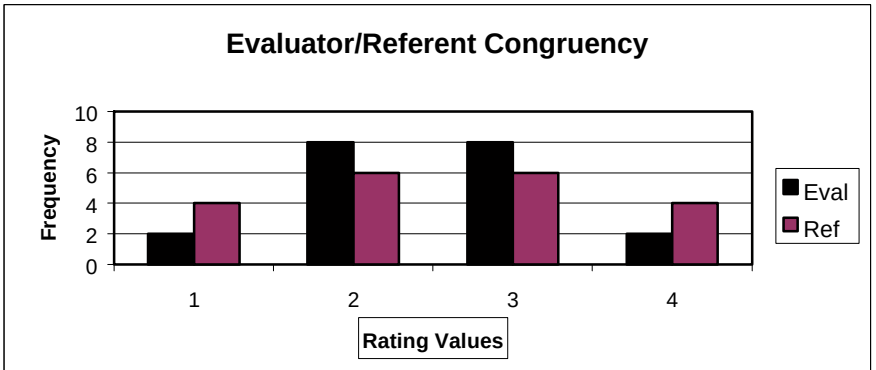

Figure 6. Congruency Between Evaluator and Referent

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.

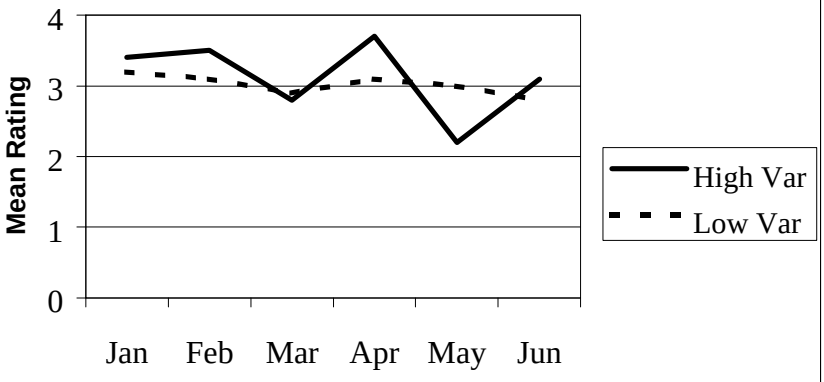

Figure 6. Control Chart for High and Low Variability Data

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		*Mean Absolute Difference from Referent*	
Individual:	*1.89*	*Management:*	*1.00*
Referent:	*2.16*	*Technical:*	*0.56*
All Participants:	*1.85*	*Overall:*	*0.79*

MyScores Correlated with Qualification Standard		*My Scores Correlated with Other Evaluators*	
Management:	*0.509*	*Management:*	*0.075*
Technical:	*0.795*	*Technical:*	*0.404*
Overall:	*0.305*	*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

Thursday, January 22, 1998

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.

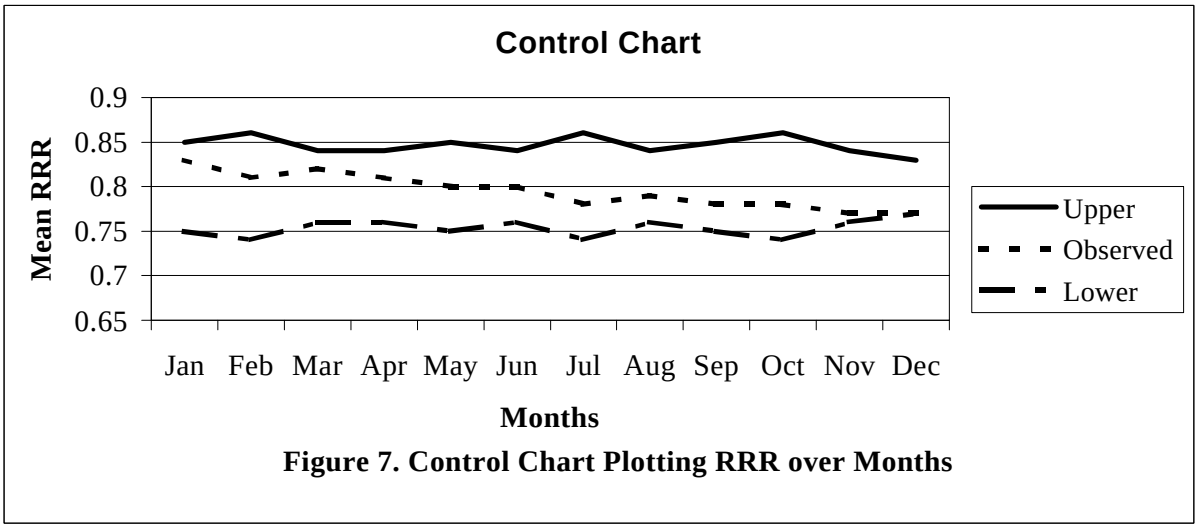

Figure 7. Control Chart Plotting RRR over Months

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