Advanced Qualification Program

Data Management Guide

Produced by the

Air Transport Association's

Data Management Focus Group

Second Release 5/12/98
TWA—RPO

This document supports the Advanced Qualification Program (AQP) and is intended for reference use only in the collection, reporting, analysis, and overall management of AQP and single-visit training proficiency data. The document was prepared by the Data Management Committee of the Air Transport Association's AQP Working Group and contains recommended information only. An individual carrier's approved AQP documentation and/or single-visit exemption takes precedence over the content of this guide. SFAR 58 and AC 120-54 contain the approved procedures for developing and administering AQP training curricula.

Acknowledgements

In 1995, the Data Management Focus Group recognized the need to develop a Guide that would enable us to share our knowledge of the many facets of data management within an AQP environment. Teams from the participating airlines considered each of these and prepared guidance, which was given to the Group Chairman, O. J. Treadway, of American Airlines, who produced a Guide that was distributed on May 29, 1996 at the annual meeting of the ATA AQP group in Minneapolis.

In 1997, the Group met to consider changes to Advisory Circular 120-54. Chapter 8, which deals with data management, was prepared by Paul Johnson of FAA National AQP office AFS 230. It contained a wealth of detail and updated information, using the Guide as one resource. The Group felt that this would serve as the basis for a new version of the Guide, and extracted the essence of the chapter that would make it parallel the input of the other groups in the Advisory Circular.

In January, 1998, the Group decided to attempt to produce the Guide for distribution at the 1998 annual meeting. With a rather short deadline, it took the efforts of many to help to prepare this Guide. Paul Johnson was, of course, of exceptional help, without which I can say that it could not have been completed. The review team included Dennis Conley, also at AFS 230, Mandy Blackmon (Alaska), Matt Humlie (Delta), Iris MacIntosh and Joy Lanzano (United), and Kevin Sliwinski (Northwest). Their hours on conference calls and very careful study of the draft have yielded a document that I believe reflects the current state of data management within AQP.

I must also recognize the help of Professor Bob Holt, of George Mason University, who has shared his passion for data and statistics with the Group through very entertaining and informative presentations at our meetings. He has also prepared inclusions for this Guide, which should serve as a starting point for those who have questions about how to handle their data.

Finally, I must thank Capt. Jack Eastman, Director of Training at TWA, for his support and for expediting the last-minute printing of this Guide.

Robert P. Odenweller (TWA)
Chairman, Data Management Focus Group
12 May 98

Table of Contents

AQP Data Management Guide
DATA MANAGEMENT CONSIDERATIONS

SECTION 1. INTRODUCTION

1. Purpose

This guide updates and replaces information first presented in the *Data Management Guide* (initial release May 29, 1996), developed by the Data Management Focus Group of the AQP Subcommittee, which is sponsored by the Air Transport Association. It provides data management guidance by discussing general data considerations and specific data collection, entry, reporting, and analysis requirements. Law and regulation in Part 121, Part 135, and SFAR 58 define these requirements. It also provides an expansion of the information provided in Advisory Circular 120-54A and incorporates data submission guidance set forth in memorandums *Streamlining Initiatives* (June 20, 1995) and *SVE/AQP Performance/Proficiency Data Submission Requirements Change* (September 19, 1996) issued by the FAA's Manager of AQP. This guide also includes findings from recent AQP grant research conducted between academia and selected AQP-participating airlines.

The principal goal of the AQP is true proficiency-based training and qualification. This proficiency-base (expressed as performance objectives) is systematically developed and maintained, then continuously validated through the collection and evaluation of empirical performance data. Data collection and analysis (data management) is, therefore, an integral part of AQP and of the SVT (Single Visit Training) exemption, which is usually implemented by carriers as a precursor to AQP.

SECTION 2. DATA MANAGEMENT

1. Definition

An AQP curriculum must include procedures for data collection that will ensure that the certificate holder provides information from its crewmembers, instructors, and evaluators that will enable the FAA to determine whether the training and evaluations are working to accomplish the overall objectives of the curriculum. In addition, it is equally important for the carrier's own personnel (fleet managers, instructors, evaluators, and curriculum development staff) to employ good data management to evaluate the effectiveness of their AQP in meeting its objectives. While airlines are expected to perform more in-depth data management, the FAA's Manager of AQP looks at higher level, global issues across all air carriers. The FAA acknowledges that data between airlines differ in many respects, and that it is neither desirable nor possible to make any meaningful cross-comparison between carriers.

Data management is required for all participants within an AQP, including crewmembers, instructors, and evaluators. It is classified into the two broad categories of *individual qualification records* and *performance/proficiency data*

2. Individual Qualification Records

An important component of any FAA approved training program is an adequate system of maintaining individual qualification records. Individual qualification records are *identifiable* records in sufficient detail on each individual who is qualifying or who has qualified under an AQP to show how and when the individual satisfied the requirements of each curriculum. These records may include demographic and work history information, as well as completion information on modules and lessons within each indoctrination, qualification, and/or continuing qualification curriculum for each crewmember, instructor, and evaluator.

Carriers may maintain a manual record keeping system based on the standard Part 121 or 135 record keeping requirements or may design a computerized record keeping system. Automated methods of collecting information for qualification record keeping are being explored at some carriers.

It is important to note that individual qualification record keeping systems are not unique to AQP. It is presumed that all carriers approved for operations have current and acceptable record keeping systems for individual qualifications. Any of these existing records and record keeping systems approved for use under traditional programs that comply with the AQP requirements and are otherwise acceptable to the FAA as meeting Part 121 and 135 requirements may be used and do not need to be duplicated for AQP.

3. Performance/Proficiency Data

Performance/proficiency data provide *deidentified* information on individual performance used in the aggregate to analyze training programs and/or groups of participants, to spot developing trends, and to identify and correct problems that may be noted. Performance data are used to determine long range trends and to support training program validation and improvement initiatives, *not* for tracking individual accomplishment. Usually, performance/proficiency data are unrelated to and separate from the individual qualification record keeping system.

The process of managing performance/proficiency data in an AQP is the primary focus of this Guide and the term data management as used throughout the remainder of this Guide refers to that process. The *data management* process consists of the four activities shown below; each of which is described in detail in subsequent sections of this guide.

- Collection of proficiency data
- Entry of the collected data into a performance/proficiency database
- Conducting statistically sound analyses on the aggregate performance data
- Reporting of the data to the appropriate managers and fleet personnel

4. Purpose

The *principal reason* for good data management at any carrier is *to establish a systematic process for quality control over pilot training and qualification.* Using data management to accomplish the four general activities attains this. General applications include:

- Using individual qualification records to validate the proficiency of crewmembers, instructors, and evaluators
- Using aggregate performance proficiency data to:
  - Validate curriculum content and/or indicate needed program changes
  - Establish expectations and performance norms for crewmembers
  - Provide a benchmark for requested and/or needed changes over time

Specific applications of the information from the data management process can be used in a variety of ways, such as:

- Provide assurances of proficiency levels
- Establish expectations and determine variations from those expectations
- Assess instructional quality
- Validate training assumptions

- Analyze effectiveness of instructors and evaluators
- Provide instructor and evaluator feedback
- Refine the training and/or measurement process
- Indicate where training changes are needed
- Validate alternative training technologies
- Provide common grounds for sharing of information between carriers
- Provide a quantitative means for CRM assessment

The carrier also may use this information to support a request for a modification of an approved AQP. For example, the carrier may request FAA approval for a three-month extension of the evaluation period in a continuing qualification curriculum (15 months instead of the normal 12 months). The carrier *must be able* to support its request with *statistically valid data* that indicate that present crewmember performance warrants the extension, and to confirm its ability to continue to collect valid data that show that performance does not degrade as a result of the extension.

5. Costs

Management of performance/proficiency data requires a commitment by the carrier of specific resources for collecting, entering, reporting, and analyzing the data. The level of resources necessary varies between carriers, but primarily is dependent on three factors.

- Extent to which the data are used (e.g., Will the data be used for process control, trend analysis, fleet comparisons, training effectiveness, cost control, compliance with regulatory requirements, requesting interval extensions, or some combination?).
- Quality of the data required for each purpose and objective (Data quality issues are discussed later in the guide.)
- Size of the carrier (amount of data to be collected).

The costs associated with data management are directly related to the number of objectives, the number of comparisons made, the amount of data collected, and the required quality of data. In all likelihood, no carrier will be willing to commit sufficient resources to collect high quality data for all purposes and objectives. However, most carriers may want to collect high quality data for a *limited set of specific objectives that are critical in the carrier's Data Analysis Plan.*

Therefore, before beginning any data collection, each carrier should determine its own objectives and data analysis plan, and then commit to the level of resources required. If necessary, a cost benefit analysis may be advisable before embarking on any serious statistical analyses.

6. Related Data Management Projects

The FAA and the aviation industry are conducting a number of related, complementary projects and data collection programs. Of particular interest to the AQP process may be the emerging data management techniques of programs that focus on the next-generation data—specifically the collection, archiving, and analytical processes of these programs. Economies of scale and simplicity of operations management may be possible by considering the impact and relationship of these programs to the overall objectives of the airline.

a. Model AQP Database

The FAA is developing a Model AQP Database for distribution and use by participating airlines as a guide in preparing their AQP at a minimal cost. The program requires automation of the Program Audit Database (PADB) and the Performance/Proficiency Database (PPDB), which can pose a prohibitive cost to smaller carriers. An FAA-developed model that can be used or imitated by any carrier will allow the FAA to be responsive in its mission to improve the safety of all air carriers without regard to financial status or size.

The first phase of development involved extensive research into the methodology that should be used to develop the curriculum for an AQP-type training program. Originally developed in Paradox, and based upon user input, the existing database was migrated to a Microsoft Access platform. The redesign of the Model AQP Database includes; migrating the previous PADB and incorporating user suggested enhancements, prototyping the PPDB and linking it to the PADB, and developing a data analysis tool for the FAA's AQP office. Carriers will be able to modify the application to meet individual needs.

b. Inter-Rater Reliability (IRR)

This study describes an innovative and comprehensive training program for improving inter-rater reliability or rater calibration for aviation instructor/evaluators using videotaped flight simulations. The training includes the

development of individual pilot and group profiles for feedback and discussion. It also offers baselines and suggests potential benchmarks as standards for rater calibration.

Pilot performance relies on systematic observation and assessment by trained raters or instructor/evaluators. Therefore, the reliability of the rater or I/E is critical to safety and flight standards. The program addresses the measurement of reliability of the I/E judgments and training that would improve inter-rater reliability.

c. Flight Operations Quality Assurance (FOQA)

FOQA embodies data collection, management, and operational insight into flight operations and associated quality control processes. As such, some of the development and implementation strategies in the FOQA program may prove helpful to AQP managers. Although the FOQA and AQP databases are completely distinct, AQP can provide tools that can support data collection and analysis as a portion of the carrier's FOQA program.

d. National Transportation Safety Board (NTSB)

Human factors receive a great deal of emphasis in an NTSB investigation into a major accident. Over the course of time, the NTSB has developed some very reliable methods for analyzing human performance data. Studying and applying some of the data management methods of the NTSB may be beneficial to some operators. This includes development of report forms, interviews, and collection processes, as well as analytical techniques that are designed to elicit detailed information on the performance of the people involved in the piloting of aircraft.

SECTION 3. GENERAL DATA CONSIDERATIONS

1. General

The process of managing data within an AQP or SVTP involves collecting, entering, reporting, and analyzing performance/proficiency information. However, before the data management process can be described in detail, it is important to discuss some general considerations regarding managing the data. This section of the Guide describes the general characteristics of data collection, the quality control aspects of data collection, some considerations for the instructor and evaluator with regard to collecting data, and the differences between management of data for an AQP versus an SVT program.

2. Characteristics of Data Collection

Certain characteristics are common to all types of data collection activities and center on three main areas:

- defining the population to be measured (who),
- defining what about the population is to be measured (what), and
- determining the methods of performing the measurements (how).

These characteristics of data collection and how they relate to the AQP/SVT data management process are described below.

3. Population for Measurement (Who)

a. General Population

The population to be measured must be defined in order to use good statistical analysis methods. Even if the population is generally known, it is still necessary to have a precise definition, because all measurements relate only to the defined population from which they were taken. For AQP, the general population for measurement is made up of crewmembers, instructors, and evaluators.

b. Sampling Units

The specific elements selected from the general population for sampling are called *sampling units*. Sampling units must be unique, easily identifiable, and selectable. A crewmember occupying a cockpit seat position is both discrete and clearly identifiable. Measurements can easily be associated with the Captain, First Officer, or Flight Engineer. Consequently, in AQP, the usual sampling units are the crewmembers scheduled for qualification and continuing qualification training and line checks within a specified reporting period, usually for one month. In addition, individual scores for each crewmember may be combined for crew-level analyses, in which case the sampling unit is the crew.

4. Items for Measurement (What)

After defining the population and the units for sampling, the *specific* items to be measured must be defined and selected. For SVT data, the measurement items primarily are the Appendix F maneuvers with conditions added. AQP data use a tailored set of Terminal and Supporting Proficiency Objectives (TPOs/SPOs) and event sets. A more detailed description of the specific items for measurement is provided later in the *Data Collection* section of this Guide.

a. Item Definition

Measured items must have clear definitions, be easily recognizable, have a clearly defined set of observable measurement parameters (qualification standards), and be given a discrete name. During training and validation, measured items normally are the individual TPOs and SPOs. However, during LOFT and LOE, these same measured items may tend naturally to flow together, such as certain items in natural combinations or in circumstances when the boundary between items becomes less distinct. When measured items are presented as a sequence of events (as in LOFT and LOE) each item must have a defined start and stop point. In addition, the instructors and evaluators must have sufficient experience so that they can recognize each measured item as it is accomplished, be able to compare the observed performance against performance standards, and assign the correct grade.

For example, consider a fairly typical event set that might consist of a CAT II approach to a landing with one engine out. This event set generally will be graded as a unit, which is acceptable as long as any conclusions that are derived are based on the *entire* event set. However, this event set also contains three discrete events, a CAT II approach, a landing, and one engine out procedures. If conclusions about the CAT II approach are desired, then items specific to the CAT II approach should be graded separately or a separate CAT II approach event should be created.

It is important to note that the definition of events for data collection may not necessarily be the same as for training purposes. For example, an LOE event set used for evaluation can be defined as a set of discretely measured items, but may also have distinct items reflecting specific training goals. Data collected on the event set should be analyzed at the level of the event set for evaluating performance and also at the level of the component training items that are included in the event set for training program feedback.

b. Item Selection

For any data collection, instructors and evaluators may have the flexibility to select from various aspects of the measured items. *While this may be good for training, it is not good for data collection.* That is because if different evaluators select different items, the variable mix of items each pilot is tested on will affect the results and make them less reliable. Therefore, for performance evaluation, the selection of items to be measured and the measurement conditions must be defined and fixed before the data collection process starts. Good definitions and consistent administration contributes to high-quality data.

For Maneuver Validation, the First Look and Fixed maneuvers are precisely defined evaluation items, while the variation in administering the Variable maneuvers aids training but at the same time makes the ratings less useful for evaluation. Particularly for First Look and Fixed maneuvers, the allowance of variations must be strictly controlled. For example, when evaluators are given a choice of NDB, VOR, and Back-Course Localizer approaches for administering a non-precision approach (NPA), the Back-Course Localizer approaches are almost never performed and the relative mix of NDB and VOR approaches given can differ among evaluators, leading to different evaluation outcomes. This must be prevented by evaluator training or direct assignment of NPA items to ensure a representative evaluation outcomes.

For an LOE, the evaluator must know the script thoroughly, because this makes grading easier and more reliable. Further, the evaluator must be trained not to depart from the script except in script variations. This ensures that each crew is receiving, as closely as possible, the same evaluation. The LOE script must be selected to be operationally relevant and developed to be a psychometrically sound evaluation (discussed later in this Guide).

Carriers with small fleets may use a single script, with all events and conditions the same for all crewmembers. The advantage is that the process is the same for all pilots and will produce the largest possible sample size for analysis. The disadvantage is that with repeated testing, crews may become familiar with the script.

Carriers with sufficiently large fleets that can support multiple samples may use multiple scripts. The script change may be as simple as pilot flying role, good day versus bad day conditions, different permutations of events, etc. Multiple scripts can provide a more complete view of crew parameters and make it more difficult for crews to become familiar with any one script. However, each permutation of scripts divides up the overall sample size into smaller

aggregates. Even the largest of fleets may not be able to use more than four script permutations. The processes used for forming reconfigurable LOE script variations are currently under study by the AQP office in conjunction with researchers and a major carrier.

Carriers should develop an LOE tracking system for each crewmember to prevent the same LOE from being used in consecutive evaluations and to track LOE variations used for each crew if such variations are implemented.

5. Methods of Measurement (How)

a. Degree of Objectivity & Subjectivity in Measurements

Measurements generally are classed as more *objective* or *subjective*, depending on the amount of human judgment involved. More objective measurements require little or no human judgment and are defined in some kind of physical unit (e.g., altitude in feet, speed in knots, heading in degrees, etc.). An example of more objective data is detailed FOQA data. Detailed objective measurements must usually be combined or transformed to arrive at a useful performance evaluation (e.g. the degree of stability on an approach). The original detailed data must be sensitive to performance differences, the combinations or transformations used on this data must be appropriate and done consistently, and the final result must be shown to be a valid indicator of pilot performance. The AQP office is currently supporting this kind of research with FOQA data. For some aspects of pilot performance, objective data may ultimately give sensitive, reliable and valid results. However, for other aspects of performance such as CRM skills, more subjective measurements will probably be necessary.

More subjective measurements require that a person evaluate the measured item and assign a score or rating. With subjective measurements each evaluator combines basic observations into a performance evaluation. Evaluators need to be trained to make a sensitive distinction between different performance levels on each item, to combine information consistently into a final assessment, and to assign an appropriate grade. The result of a subjective evaluation process must also be assessed for measurement validity. *Recurrent training in evaluation and data collection procedures for evaluators is just as important as recurrent training in flying procedures for the crew.* This procedure is similar to the regular checks performed on the calibration of devices that record more objective data, such as altimeters or airspeed indicators.

In most instances, AQP and SVT data are measured subjectively. The data consist of repeat counts and event ratings that are dependent on the subjective judgment of the instructor/evaluator. Therefore, it is essential to good data collection that outside biases do not influence the judgment of the evaluator.

b. Samples

A sample is a subset of the population and/or performance events. Conducting evaluations on a sample allows the parameters of the general population to be *estimated* from the sample results. As population parameters must be estimated, it is not possible to determine the *exact* parameters of the *entire* population. However, the analyst can determine how close these estimates are with a known level of certainty. The certainty of these estimates depends on the sample size and amount of variation in the item or index.

The AQP continuing qualification and SVT recurrent training programs require that *all* crewmembers be evaluated each period and/or cycle. Consequently, AQP and SVT data have *some* characteristics of a census in that some data will be collected on almost all sampling units in a fleet (crewmembers, instructors, and evaluators). However, it is different from a census in that data are not taken from all sampling units at once but rather monthly samples spread out over the year. In AQP continuing qualification curriculum and SVT, a fraction of the crewmember population is evaluated each month on a subset of events, such that all crewmembers are evaluated within the period (usually 12 months). Each monthly subset can be considered a *sample* from the parent population. Since the monthly sample size is fixed by fleet size and by the number of fleet pilots requiring evaluation for currency, the estimates from a single month cannot be improved by increasing the monthly sample size.

c. Sampling Methods

The two common sampling methods are *random* and *representative*. In a random sample, all sampling units in the population are *equally likely* to be selected and there are NO systemic criteria associated with the selection method. This makes the sample a true reflection from the parent population, only smaller. In a representative sample, the sampling units are selected by some systematic criteria other than equal likelihood.

The pilot selection method for AQP and SVT is the set of crewmembers that need to receive continuing qualification or recurrent training this month. Such a sample is considered to be *representative* of the parent population, but is not truly random. For example, due to intense waves of hiring pilots for a fleet, a *bolus* or clustering of pilots with certain characteristics may predominate in a given month. This bias in the monthly sample could conceivably influence the results. However, if there is no strong, systematic shift in pilot samples across sampling periods, the samples may be representative enough for statistical use.

A second bias to be avoided in the selection method is any systematic assignment of evaluators to pilots and crews. Evaluators must NOT be allowed to choose the crews they evaluate. Further, the scheduling of evaluators and crews for an assessment should be examined to ensure that it depends only on availability and NOT on any other systematic variable. For this purpose, *computerized assignment of evaluators to crews is clearly preferable to clerical assignment.*

A third bias to avoid is any systematic shift in the measured items over months—measured items must be the same for all monthly samples. Any systematic shift that would result in a change in *which* items are measured in certain months (e.g. a shift in the relative proportion of NDB vs. VOR Non-Precision Approaches) can bias the monthly results.

d. Aggregating or Dividing Samples

A characteristic of random samples is that any division or aggregation of a random sample is also a random sample. We make use of this property when we divide the sample by aircraft, crewmember position, maneuver, etc. to obtain more detailed statistical information. The sample can be subdivided safely into various subclasses as long as the resulting subsamples satisfy the *minimum* sample size restriction for doing statistical analyses.

It is also possible to combine two or more monthly samples to create a sufficiently large sample to measure interactions among critical determinants of performance. For example, the interaction of aircraft by seat (Captain, First Officer) by maneuver on maneuver validation may require aggregation over several months to ensure sufficient data. Aggregating monthly samples into yearly results will give a sample that is very close to a complete census for the fleet and is even more representative of the fleet population than any monthly sample. Therefore, once-a-year analyses may be performed at any natural juncture of the data collection process, such as the switch to a new LOE for the fleet.

SECTION 4. COLLECTING QUALITY DATA

1. General

The collection of quality data for an accurate representation of the population requires capturing a high quality of data and a low level of errors in the data. Sample sizes are discussed in the *Data Analysis* section of this Guide.

2. Data Collection vs. Evaluation

Several conflicts exist between the requirements for AQP/SVT evaluation data collection and crewmember training. These conflicts are centered on the rigidly controlled environment required for collecting good quality evaluation data and the more loosely structured environment required for good quality crewmember training. Some events like LOFT clearly emphasize training at the expense of evaluation. Other events like the LOE emphasize evaluation during the critical simulator portion while training is relegated to prebriefs, debrief, SPOT checks, and so forth. Carriers should always keep in mind that, while good quality data is important, *the primary purpose of AQP and SVT is better training*. **Never let the data collection process overwhelm the training objective.** If push comes to shove, the training goal should dominate and modifications in the evaluation procedure should be noted in the data collection process (e.g. noted on the form). In this way, training is not jeopardized while at the same time the evaluation data can be "cleaned" of non-standard evaluation sessions so that the descriptive and inferential statistics are accurate.

A good training program should permit some instructor discretion. It may also allow the training program to be tailored individually to the crew being trained. The training scenarios and syllabi are commonly used as guides for what must be accomplished during the training period, rather than as rigid scripts. However, this kind of discretion introduces errors into the data that may, at times, make them less reliable. This may occur for training evaluations such as LOFT. However, it is also the case that good training depends on accurate evaluation of a particular pilot's strengths and weaknesses in performance. Therefore, accurate evaluation is also an important part of training sessions such as LOFT. If evaluation quality is basically preserved, data can be captured from the more relaxed training environment and still provide useful information.

High quality evaluation data can be collected for a specific *subset* of all data, such as the First Look section of Maneuver Validation and the simulator portion of the LOE. For these subsets, the evaluation goal should have the *same priority* as training. This limited set of high quality data should be collected with sufficient attention to detail such that performance will be evaluated reliably.

In fact, the initial analysis by the FAA of SVT data from several carriers suggests that data from a subset of more carefully controlled events will produce better statistical results than data collected during a more general training session. The First Look data usually have only a limited set of specifically defined maneuvers and conditions that are administered to both the Pilot in Command (PIC) and the Second in Command (SIC). A statistical review of the data has shown a very marked distinction in the quality of these maneuvers as compared with data gathered during the more general training sessions.

3. High Quality Data Attributes

Four primary attributes determine the suitability of data for analysis that will yield meaningful results. The greater the attention to these attributes, the higher the quality and the more useful the collected data become. The main determinant in the quality of data is in the preservation of these data attributes.

a. Sensitivity

Sensitivity means that small gradations or variations in the parameter being measured, such as maintaining heading in a maneuver, are reflected in some variation in the measurement, such as the scale rating of the pilot performance. For a single item, a multiple-point scale will allow more sensitivity in measurement than pass/fail grading. Alternatively, using multiple items to measure individual components of performance and then combining those scores into a composite score or index could increase sensitivity.

b. Reliability

Reliability means a lack of random error in the final measurement, which is usually indicated by consistency among items or stability in the measurements over time. Reliability is affected by both sampling and non-sampling errors. Low Inter-Rater Reliability is an example of a problem that could cause performance estimates to be unreliable.

c. Validity

Validity means that the data accurately measure what they are intended to measure. Data validity depends partly on having adequate sensitivity and reliability of the measures, but even with adequate sensitivity and reliability the data may be measuring something other than what is intended. For example, a poorly worded grading scale may encourage grading on the evaluator's liking of a pilot's style rather than the pilot's objective performance. Data validity is weakened or reduced by vague data definitions, insufficient evaluator training, casual data collection methods, operator discretion, etc.

d. Good Data Collection: Standardization and Completeness

Standardization means the data collection procedures are uniformly followed. Consistency ensures that data differences do not come from procedural differences. The standardization attribute is very sensitive to non-sampling errors associated with instructor/evaluator training and the stability of established procedures. Completeness means that all the expected data elements, records, observations, etc. are present. Completeness is affected primarily by non-sampling errors. Common causes of incomplete data are: failing to record a maneuver, recording only part of the required information, losing session records before entering data, data collection documents becoming separated, inability to read/scan the data collection form, corrupted files, etc.

Data collection must satisfy the following criteria for the measurement to be of high quality:

- The events to be measured must be pre-selected by some process. This could be a fixed set, random, representative, combination, or other statistically defensible method. Selection of events to be validated/evaluated *can not* be at instructor discretion.
- The measured events must be well defined with discrete performance, conditions, and standards.
- The events to be measured must be separate and distinct. Coupled events (e.g., Non-Precision Approach to a Missed Approach) must be rated and analyzed as one contiguous event, or the individual components must be rated separately if they are to be analyzed separately.
- The instructors/evaluators must be required to make an explicit grading of all ratings, repeats, reason codes, etc. Rating by exception implies an answer and has the potential for omissions.
- Instructors and evaluators who accomplish the data collection must be trained to explicit standards that are well defined.
- Data on the measured events must be collected in a session that has been set aside discretely for validation purposes. Training should not be accomplished until the validation data collection is complete. (e.g., First Look should not be intermixed with training, etc.) Note that the "validation session" may be the initial part of a regularly scheduled session, rather than one that stands alone.
- Devices (flight simulators) used for event measuring sessions must be of similar fidelity. For data purposes, similar devices are considered to be levels 6/7, A/B, or C/D. For example, if data from

some First Look sessions are captured in a level C simulator, then all First Look data should be taken in a level C or D simulator and not mixed with data taken from a lower level device.

- The rating scale must be well defined, observable, and consistent. (i.e., Two instructors observing the same performance should be able to mark the same rating with the same reason codes.)
- The volume of data to be collected should be limited to avoid instructor overload.
- The procedures for handling and entering the collected data should be well defined to avoid loss of information.
- No deviations from prescribed procedures should be permitted.

As can be seen from the above list of requirements, data collection for inferential statistics is serious business and attention must be paid to the details.

4. Use of High- or Low-Quality Data

Collected data can be used to estimate aspects of the population such as an average (*descriptive* uses) or to answer specific statistical questions (*inferential* uses). Data that have been collected with emphasis on preserving the above four data attributes can be used to make estimates of population parameters, such as average performance, at relatively higher levels of confidence. Similarly, there is more confidence in answering statistical questions with these high-quality data. Data that have not been collected with regard to preserving these attributes give fuzzier and possibly misleading results when used for descriptive or inferential purposes. Clearly, higher-quality data are preferable to lower-quality data for all purposes.

The primary difference between high and low quality data from a management perspective is in the cost and complexity of the data collection. High quality data are more rigorous, costly, and time consuming to collect. Consequently, high quality data collection may occur only where necessary on a limited subset of events and/or validation sessions. Clearly, evaluations such as First Look maneuvers and LOEs require high quality data collection. Lower quality data collection is acceptable for non-evaluative measurements (e.g. LOFT), and measurements that are simple adjuncts to the training process. In addition, carriers should be guided by the following uses for higher and lower quality data.

- Inferential quality data ***should*** *be used* in AQP for:
- ➢ Extending the continuing qualification cycle
- ➢ Reducing or eliminating individual events from training or checking requirements
- Descriptive quality data ***may*** *be used* in AQP/SVT for:
- ➢ Reducing or extending the curriculum footprint
- ➢ Changing the set of fixed and/or variable maneuvers (or events)
- ➢ Validating currency items

5. Non-Sampling Errors

Non-sampling errors are *errors* that are embedded in the data collection and analytical process rather than the process of sampling pilots. These errors are associated with the process or procedures used to collect, process, and analyze data. These errors may be systematic or non-systematic. Systematic errors are ones that follow a pattern or are controlled by a function that may bias the observations in a specific direction. Non-systematic errors are ones that occur at random.

Systematic errors will bias any statistical analysis and produce misleading results. For example, if evaluators shift their use of a rating scale so that they use higher grades for the same performance, the data analysis may indicate a significant increase in performance, which would be highly misleading. Systematic errors are not affected by sample size and they cannot be corrected by repeated sampling. *Systematic errors must be recognized and corrected.*

Non-systematic errors add "noise" to the data. If this "noise" is not controlled, inferential analyses can give indefinite results in which the answers to questions can be hidden. For descriptive analyses, non-systematic errors can cause more unpredictable and erratic estimates. *Non-systematic errors should be minimized or avoided.*

The causes of non-sampling errors are many and varied, but generally fall into the relatively broad categories of data definition, procedure, measurement, and data handling. Examples of non-sampling errors are:

- Improperly reporting the required data collection items (e.g., Not reporting incidences of repeats, missing reason codes, not recording items, recording items incorrectly, improperly marked bubble sheets that cause the scanner to miss the intended mark, recording the last rather than the first occurrence of a measured item during a First Look session, etc.)
- Data handling errors (e.g., losing forms, separating simulator day 1 and day 2 forms, data entry or coding errors, etc.). There may always be some invalid entries, but data entry procedures can be constructed in a way that will screen the input for validity and help to reduce these errors.
- Allowing events for measurement to be selected at the discretion of the instructor/evaluator (e.g., requiring a Non-Precision Approach to be accomplished, but allowing the instructor to select *which* Non-Precision Approach)
- Data definition errors (e.g., Overlapping or inconsistent data definitions are caused by ambiguities in describing the data to be collected. The items being measured must be clearly identified, otherwise the data collector, data analyst, and data user may be considering dissimilar items. In effect, they talk past each other. For example, consider two maneuvers defined as Non-Precision Approach to a Missed Approach and Non-Precision Approach to a Landing. Depending on the viewer, these could be considered as two, three, or four different maneuvers—two Non-Precision Approaches, one Missed Approach, and/or one Landing. If the component parts are to be considered a single maneuver and a repeat is required, is it because of the Approach, the Missed Approach, or the Landing?)
- Evaluator grading errors (e.g., AQP and SVT measurements are judgment calls by instructors and evaluators. Subjective measurements of performance, poorly defined measurement standards, and lack of training on the use of subjective grading criteria can cause such errors.)
- Lack of standardization among instructors and evaluators (e.g., The changing mix of instructors and evaluators may affect opportunity to provide standardization training for instructors and evaluators, which in turn can cause the performance assessment process to become somewhat like having a "rubber ruler.") (Refer to the discussion on *Instructor/Evaluator Considerations* later in the Guide.)
- Data collection procedures and policies that favor one response over another (e.g., The use of exception reporting and default responses.) This virtually guarantees that adjacent ratings will be *under-reported.*
- Problems with data collection forms (e.g., poor or complex form design, using incorrect or obsolete forms, etc.)
- Reporter fatigue (e.g., Long forms, numerous items, and detailed responses can cause the instructors and evaluators to rush through the forms without proper consideration.)
- Errors in the computational algorithms used for sorting, reporting, and analyzing the data.

SECTION 5. SVT VS. AQP DATA COLLECTION

1. General

AQP data collection can be split into two broad phases. The first phase is data collection during the transition from the traditional recurrent training program to the actual AQP. The second phase is data collection associated with the approved AQP itself.

Collecting SVT data is essentially a data collection process superimposed over the traditional *recurrent* training and line check program. Data are collected for all maneuvers or items performed or checked and consist of grades and/or repeats to proficiency with associated reason codes. AQP data collection, on the other hand, demands more rigorous attention and is required in *all* AQP curricula. However, good data management is required for both the AQP and the SVT.

The differences and similarities between data management for the AQP and SVT are described in the following table. Keep in mind, however, that the AQP data management requirements *replace* the SVT requirements for a specific fleet as soon as the AQP is approved for that fleet. Therefore, this Guide focuses primarily on AQP data management.

2. SVT/AQP Differences

SVT/AQP Differences		
Requirement	**SVT**	**AQP**
Data Management Procedures	Similar for all carriers and are explicitly defined by FAA in each carrier's exemption.	Established by SFAR 58 and recommended procedures established in this Guide. Ultimately, defined and agreed on between the individual carrier and FAA Manager of AQP, and documented in carriers AQP plans.
Measured Events	Defined by maneuvers.	Derived from tasks/subtasks in front-end analysis and TPOs/SPOs in curriculum design.
Performance, Conditions, and Standards	Based on published practical test standards. Does not allow for variation from practical test standards for licensing purposes.	Allow deviation from practical test guide and are based on the Qual Standards document.
Reason Codes	Requires reason codes to be reported for substandard performance on measured events. No explicit requirement to report CRM data to FAA.	Requires identification of proficiency objective standards that were not met, including CRM standards.
Data Reporting	Required only for crewmembers in recurrent training (continuing qualification).	Required for all participants (crewmembers, instructors, and evaluators) in all curricula (qualification, and continuing qualification).
Instructor/Evaluator Training	Does not require instructor and evaluator training beyond the traditional regulations.	Requires that instructor and evaluator curricula (indoctrination, qualification, and continuing qualification) are designed and the application of measurement procedures is incorporated.
First Look	Requirements highly abbreviated (limited to a core set of maneuvers) and proficiency data may be obtained during normal training rather than in an explicitly convened first look session. There is no requirement to assess proficiency on currency items or to use first look data as a basis for considering the interval between training and checking.	First Look data are defined by criticality and currency analysis and may be collected in any CQP simulator period.
LOFT Data	Usually contains recurrent LOFT, but no LOE.	Contains both LOFT and LOE. Data are not required to be collected or reported for the LOFT.
Line Check Data	Obtained only on PIC, rather, than the entire crew, and there is no requirement to assess currency item proficiency during line checks.	Required on a crew basis for both proficiency validation and currency assessment.
Fleet Data	Required on all fleets.	Required only in fleets of aircraft that have transitioned to AQP.
Data Uses	Not used for program validation purposes, but as a control tool to assure continuing safety of SVT.	Employed to validate AQP approach to training system development, to determine future disposition of CRM for pass/fail purposes, and to provide empirical basis for rule making on future disposition of AQP.
LOE	Not required.	Required for AQP Qual and CQP overall grade and event set grade.
Validation/Evaluation	Failures, repeat counts, and reason codes.	Comments or reason codes, overall period grades, comments or reason codes for unsatisfactory performance.

3. SVT/AQP Similarities

SVT/AQP Similarities	
Requirement	**SVT and AQP**
Continuing FAA Approval	Continuing FAA approval for SVT is contingent on the accomplishment of reasonable progress towards transitioning all carrier fleets to an AQP.
Reporting of Electronic Data to FAA	Monthly reporting (but not greater than 60 days following month of collection) of data in digital electronic format to the FAA Manager of AQP is required.
Airplane & Duty Position Reporting	Data reporting is aircraft and duty position specific. While SVT reports maneuvers, their analogues in AQP are the terminal and supporting proficiency objectives.
Anonymous Data	Data are de-identified, but are reported with a given anonymous crew constituting one record.
Data Collection Sessions	Data are reported from first look, maneuver evaluation, proficiency evaluation (either SVT proficiency or AQP LOE), oral, and line check activities.
Data Format	Data reporting format is individually negotiable between the carrier and the FAA Manager of AQP.
Trend Analysis	Global trend analysis is employed on all data. The FAA Manager of AQP provides individual carriers with feedback on the implications of their data with regard to training program effectiveness. Carriers also are responsible for conducting their own analyses of such data and initiating program modifications as necessary to address negative trends.

SECTION 6. DETAILS OF AQP DATA COLLECTION

1. General

Data are collected for proficiency and enabling objectives for each individual crewmember, instructor, and evaluator participating in an approved AQP curriculum. The data are collected at strategic points within each curriculum and these points of collection are selected to provide the most accurate information for the desired results. This section of the Guide describes the data collection process, i.e., what data to collect and at what points in the curriculum is collection most advantageous. Keep in mind that all performance data collected on each objective must be relative to the applicable qualification standards defined for the training and evaluation activities.

2. Data to be Collected

An Advanced Qualification Program is made up of terminal and supporting proficiency objectives (TPOs and SPOs) and enabling objectives (EOs). The TPOs and SPOs typically are defined as *skills* or *behaviors* that can be trained and *observed*, usually in some sort of training device. TPOs and SPOs are described in the qualification standards document, which specifies the duty position(s), performances, conditions, standards, and applicable document references on which basis the proficiency of an individual is assessed.

Individual TPOs and SPOs sometimes are referred to as “events.” Events usually are the defined procedures and maneuvers in training and/or validation sessions. Grouping of the TPOs and/or SPOs into logical units for instruction and evaluation is referred to as an “event set.” Event sets normally are used within the context of LOFT and LOE.

EOs are the lowest level to which the task hierarchy is divided and typically are defined as *knowledge* and some *lesser order skills*. EOs usually are learned in the classroom by use of a part-task training device or through self-study, and initially validated via some form of *knowledge testing*. Subsequent evaluation of EOs may involve practice and observation in a device to the desired proficiency level.

For indoctrination and initial qualification, data on all relevant TPOs, SPOs, and EOs is required at one or more points in an AQP curriculum. For CQP, only a subset of TPOs and SPOs can be assessed at each yearly cycle. However, the set of TPOs and SPOs assessed each year should be systematically changed so that all relevant TPOs and SPOs are covered during a multi-year evaluation cycle. One approach to collecting these data is illustrated in the

table below. In addition, data collection usually involves gathering ancillary information for data grouping and analysis purposes, such as dates, session types, instructor/evaluator information, crew position, region, etc.

Event Set	Training Assessment of flying scenarios	LOE Event Set Grades
TPO	Training Assessment of general Technical or CRM knowledge	LOE Technical or CRM Task Grades
SPO	Training Assessment of specific Technical or CRM knowledge	LOE Technical or CRM Skills Grades
EO	Training Assessment of specific procedural steps and detailed knowledge	LOE Observable Behaviors Grades

3. Performance Data (TPO/SPO/Event Set)

Performance data are usually collected during training, validation, or evaluation in some sort of device (CPT, FTD, simulator, etc.). The recommended specific measurable items to be collected with regard to performance data are:

- Aircraft Type
- Curriculum Type
- Type of Session or Session Identifier
- Date and/or Time Information
- Instructor/Evaluator Identifier
- Crew Position (Captain, First Officer, Flight Engineer, instructor, and evaluator)
- Pilot Flying/Pilot Not Flying Identifier
- Geographic/Regional Area of Evaluation (Applicable to line check only)
- Multi-Segment Required (Applicable to line check only)
- Use of Seat Substitutes (Position used)
- Overall Session Rating
- Referred for Additional Training (Yes/No)
- Eliminated From Training (Yes/No)
- Event Identifier (TPO, SPO, or event set alphanumeric identifier and/or name)
- Event Rating (For each TPO, SPO, or event set observed)
- Repetitions to Proficiency (For each TPO, SPO, or event set, as applicable)
- Reason Code(s) for Substandard Event Ratings *(Following are examples only)*
- ➢ Automation (FMS, GPS, FGS operation)
- ➢ Execution (Technical performance/proficiency)
- ➢ System Knowledge
- ➢ Procedural Knowledge or Compliance (Policy, Procedural, Regulatory)
- Crew Performance Indicators:
- Communication

- Situation Awareness
- Workload Management
- Teamwork

Note the following two points with regard to the above measurable items: First, in order to facilitate curriculum maintenance and correction, the item identifiers should be similar (if not identical) throughout the analysis, design, implementation, and data collection stages of AQP. At a very minimum, the assessed performance items *must* be precisely linked to the appropriate AQP curriculum segments in the Program Audit Database so that results of analyzing the performance data can directly feed back to recommended changes in the AQP curriculum. The identifiers for the TPOs, SPOs, and EOs should correspond to the identifiers used on the respective tasks, subtasks, and elements from which the events are derived. Second, reason codes generally are carrier specific. The codes shown above are for example only.

4. Enabling Data (EO)

Traditionally, the knowledge required for enabling objectives is considered to be a part of "ground school." Typically, most carriers have graded ground school objectives on an overall pass/fail or complete/incomplete basis via computer, written, or oral exams. Because of this, there have been little data collected on a per EO or question basis.

However, to validate the "ground school" training in an AQP and improve the curricula, exams should be administered covering the knowledge required for specific enabling objectives and enabling performance data (where feasible) should be collected and analyzed. Whether these exams are administered via computer, written, or orally, all questions should be objective, consistent between students, cover only the defined enabling objectives, and follow the same rules for quality data collection as described previously for the performance data. Otherwise, any EO data collected may be unreliable for curriculum validation and/or correction. For example, oral exams with no consistent structure and content across pilot trainees would yield unreliable or unusable data for curriculum change.

Note that in any curriculum a single enabling objective may be made up of multiple instructional points and likewise require a set of one or more questions for validation of required knowledge for the EO. Therefore, data should be collected on *each individual question* within the EO question set to allow proper test item analysis, as well as validation of the overall enabling objective. The recommended specific measurable items to be collected with regard to knowledge required for enabling objectives are as follows:

- Aircraft type
- Curriculum type
- Type of session
- Date and/or Time Information
- Instructor Identifier
- Evaluator Identifier (use special codes for computer-based evaluation or written test)
- Crew position (Captain, First Officer, and Flight Engineer)
- Overall test completion data and scores
- EO Identifier (alphanumeric EO identifier attached to each question within the EO question set)
- Individual question results for test item analysis

It also should be noted that in addition to the above scheme for measuring performance of enabling objectives, some carriers might use EOs more like "reason codes" attached to performance of higher level TPOs and SPOs. When used in this manner, the qualification standards that define TPO/SPO performance also contain the specific enabling

objectives relating to each TPO and SPO. During performance training and validation, any TPO/SPO rated substandard must also have an indication of the factor that contributed to the substandard performance.

5. Demographic Information

While not required, some carriers may desire to collect various kinds of demographic data on the crewmembers, instructors, and/or evaluators to aid in overall analysis of the training programs. If collected, demographic data must be tailored to the needs of each carrier, but may include such items as crewmember background (military, civil, foreign, etc.), total flight time, total time in type, recency of experience, experience in seat position, automation experience, "glass cockpit" experience, etc.

6. Crewmember Identification

Although de-identified data are sent to the FAA, it would behoove the carrier to affiliate performance/proficiency data to individuals and the instructors/evaluators (e.g. using a randomly assigned Personal Identification Number). This individual-data affiliation would provide the carrier with increased data analysis and quality control capability.

SECTION 7. MEASUREMENT CODES

The applicant will identify the rating methodology that will be used to grade the performance of the proficiency objectives against the qualification standards. Typically, the measurement codes associated with performance events are ratings, repeat counts and reason codes or skill categories. Ratings are used to define different quality levels of performance. Evaluators must be well-trained in order for the process of rating to be less subjective and idiosyncratic. Rating codes usually are carrier specific and the FAA requires that something more sensitive to performance differences than a binary code be used (i.e., some rating method that provides more performance differentiation than "pass/fail").

Each carrier should ensure that the grades established on the rating scale are clearly defined, meaningful to the I/Es, and easily used for performance assessment. Each point in the rating scale must also be communicated to the FAA as de-identified data that is submitted along with reason codes and repeats. Although consistency among fleets and across different types of evaluations (Line Check, Maneuver Validation, Online Evaluation, and LOE) is important and generally desirable, rating scales may be slightly different when used for different purposes such as training vs. evaluation. Listed below is an example of a rating scale that discriminates among performance levels:

First Look, Maneuvers Validation, LOE, or On-line Evaluation

GRADE		CRITERIA
(1)	Unsatisfactory	Deviations from the Qualification Standards exceed prescribed limits and are not corrected in a timely manner commensurate with safety.
(2)	Satisfactory	Deviations from the prescribed Qualification Standards occur but they are recognized and corrected in a timely manner.
(3)	Standard	No deviations occur from the prescribed Qualification Standards.
(4)	Excellent	Performance remains well within the prescribed Qualification Standards and management skills are exemplary

This example should not be taken as limiting possible scale intervals to a four-point scale. With appropriate scale construction and instructor/evaluator training, carriers may elect to define other scales which maximize the quality (sensitivity, reliability, validity) of the collected data.

SECTION 8. INSTRUCTOR/EVALUATOR CONSIDERATIONS

1. General

Instructors and evaluators are the first link in the data collection process. They are responsible for training the crewmembers, administering tests and evaluations, operating the equipment, and recording the results. To obtain accurate performance/proficiency data and still preserve the effectiveness of the training and/or evaluation sessions, careful consideration is required by training managers and data analysts.

2. Training

Since much of the evaluation, event rating, and data collection process is *subjective*, it is imperative that instructors and evaluators understand the evaluation and data collection process so that events can be properly scored. Therefore, each carrier must have a thorough training program for their instructors and evaluators. AQP requires that each instructor and evaluator receives qualification training (or differences training for existing personnel), continuing qualification training on a periodic basis, and requalification training, as necessary, after an absence. The importance of including evaluation standards and data collection procedures in all of these instructor/evaluator-training programs cannot be overemphasized.

Training should be as all-inclusive as possible. It should include a review of the definition of items to be measured, the standards for measurement, the rating definitions, the rating process, how to rate coupled events, the collection of ancillary data on events (i.e., reason codes, trials to proficiency, etc.), use of the data collection instruments, and the data entry procedures, as applicable. It is vitally important during these instructor/evaluator training sessions that, in addition to presenting the material, *practice and feedback* is included in the rating of events and the overall data collection process.

Evaluations are subjective, and, without a periodic review of the data collection process, each instructor/evaluators concept of proper rating tends to drift away from standard and become individually biased. Therefore, the continuing training program should include a periodic group standardization session for the instructor and evaluators so that their grading standards and data collection practices can be similar. Inter-Rater Reliability (IRR) training is one method to help calibrate and standardize instructors and evaluators.

3. Rotations

Most carriers incorporate some form of instructor/evaluator rotation, i.e., periodically cycling between instructor/evaluator duties and other carrier responsibilities, such as flying the line. While rotation may be good for training, it may have a negative impact on the quality of data collection due to potential problems with instructor standardization. The effect of rotation may be to create multiple rulers by which the crewmembers are rated. Each ruler is the specific set of instructors and evaluators who are rating for the particular rotation period.

No two instructors see events exactly the same way or rate in exactly the same manner; some rate hard and some easy. These differences should be identified and reduced by IRR training, but it may not be possible to eliminate them completely. Each carrier should set a limit on the extent of systematic instructor differences that it will tolerate and check these differences during the regular data analysis process. If the pool of instructors remains constant, the relative number of hard and easy graders tends to stabilize over time, and the collected data become stable. However, if the mix of instructors changes from month to month, the changing mix may cause additional variability in the ratings and data collection process, which could contribute to unreliable data.

It should be noted that the effects of rotation on AQP data collection are not yet completely known. However, the FAA's Manager of AQP commissioned a research study to investigate Inter-Rater Reliability and the effects of improving crew assessments. In the interim, carriers may want to try to minimize the rotation requirements for those instructors and evaluators involved in the early stages of AQP. If rotation is necessary, an adequate training program

is the best way to mitigate the effects of rotation. Each time an instructor or evaluator is rotated back into the training program, some refresher training in event rating and data collection may be in order. The amount of refresher training is dependent on several factors, such as the instructor's time away, the degree of standardization in the training programs, the overall stability of the data, etc.

4. Instructor Limits on Data Collection

In addition to their previous duties as instructors and evaluators in a traditional training program, under AQP the instructors and evaluators are now responsible for collecting the performance/proficiency data. A balance between collecting the minimum data necessary for a meaningful evaluation and collecting highly detailed information about performance of each maneuver requires careful consideration by those managing the training program. Recording too little may result in data that are relatively useless in validating training and formulating corrective action. Recording too much data could turn the instructor/evaluator into a very busy recorder of events and significantly degrade the quality of instruction and evaluation.

Obviously, there is a limit to the number of data points that an instructor/evaluator can collect in an individual training session before efficiency is affected. Current research sponsored by the FAA Manager of AQP indicates a typical instructor/evaluator is able to enact and intensely observe 10 to 12 maneuvers or five to eight event sets per pilot in a two-hour simulator session. For each event set, the I/E should complete a basic set of ratings of overall performance for each pilot plus repeats required and reason codes if necessary. Important ancillary data for the LOE can range from rating specific observable behaviors to rating technical or CRM skills or tasks.

The maximum number of good-quality judgments an I/E can make during a simulator session depends on the training of the I/E, the user-friendliness of the rating form, and the flexibility in the rating process. At optimum, I/Es are carefully trained, the form is well-designed taking human factors into account, and the rating process is designed to allow I/Es to complete ratings in low-workload periods during the simulator session. Under these conditions, current research indicates that an I/E can reliably evaluate 4-6 observable behaviors and 3-6 specific skills or tasks in addition to the basic ratings on each event set. This research indicates that the entire data collection session an I/E can reliably observe 100 to 150 elements, but may at the time make note only of those that are outside standard expectations. Exceeding 150 ratings in a simulator session may result in the I/E using response sets rather than truly grading each item or responding in a stereotypic manner by giving, for example, all "standard" ratings. This type of error in the rating process due to I/E overload must be avoided in LOEs, maneuver validations, and line checks.

SECTION 9. ACTIVITIES REQUIRING DATA COLLECTION

1. General

Data collection requirements for the AQP vary with the phase of AQP development, the curriculum (indoctrination, qualification, or continuing qualification), the type of curriculum activity (training, validation, evaluation, etc.), the type of participant (crewmember, instructor, or evaluator), and the overall management objectives for data use. The data collection requirements for the various categories are described below.

2. Phase III (Implementation)

Data collection requirements during the implementation phase focus on the instructor/evaluator and student perceptions during small group tryouts, if used.

a. Existing Instructor/Evaluator Training

Existing personnel must receive knowledge training regarding the differences in their job roles between the traditional training curriculum and the AQP curriculum about to be implemented. Enabling objective data should be collected on at least the overall test completion and individual question scores.

Carriers should record evaluator-identifying information on gradesheets for quality control. This will also permit the carrier to conduct its own internal "Inter-Rater Reliability." The FAA inspector's name or identification number should be recorded on gradesheets whenever an evaluation is observed.

b. New Instructor/Evaluator Training

New instructors and evaluators must receive knowledge and performance training as instructors/evaluators in the AQP curriculum. Objective data should be collected appropriate to the respective curriculum segments involved (ground school, flight procedures, maneuvers, flight operations, line evaluation, etc.) in the instructor/evaluator AQP curriculum.

c. Small Group Tryouts

During small group tryouts of the pilot training curriculum, data collection should principally entail the written comments of skilled observers for various curriculum segments, as well as student critiques of the course.

3. Phases IV & V (Initial & Continuing Operations)

Performance data on TPOs, selected SPOs, and event sets, as well as enabling data should be collected on all crewmembers, instructors, and evaluators, as appropriate, beginning in phase IV and continuing throughout the life of the AQP. As outlined in Section 13, performance/proficiency data will be submitted to the Manager of AQP on a monthly basis and no later than 60 days following the month of collection. Principal activities during which performance and enabling data are collected are described below.

a. Training

Training data can indicate whether the knowledge or proficiency of an individual is at an appropriate level for the curriculum day/module. On an aggregate basis, training data permit a determination of whether a curriculum segment delivers appropriately qualified pilot trainees to the next curriculum segments. Of particular importance, training to proficiency may be substituted for the formal validation sessions on many items in an AQP, both for qualification and continuing qualification. Collecting and analyzing training to proficiency data may enable regulatory approval of such an approach in lieu of the formal validations. In continuing qualification, these data supplement first look and maneuver validation data by tracking how much training is required to re-establish proficiency on tasks not initially performed to standards.

Training data usually are collected on a daily basis and are applicable to both qualification and continuing qualification. During *qualification* training, data should be collected for *both knowledge and performance* information, whereas, *only performance* data usually are collected during *continuing qualification*. Proficiency data should include the module where items are introduced, repetitions to proficiency, and the module where proficiency is achieved. Knowledge measurement data should include individual test item information.

b. Validation

Validation data are the same as training data, but also add individual maneuver ratings, reason codes on substandard events (for diagnostic, not jeopardy purposes), and overall session ratings. Data fields should also be present to indicate a referral for additional training and/or elimination from training, as necessary.

c. Line Operational Evaluation (LOE)

LOE data are *performance* based and are applicable to both qualification and continuing qualification. LOE data collection must include at least an LOE session identifier code, the overall LOE session rating, individual event set identifiers with ratings, and data fields to indicate a referral for additional training and/or elimination from training, as necessary. LOE data collection also must include a performance rating on each TPO/SPO that makes up each event set, along with reason codes for substandard performance that reflect meaningful performance categories for a given carrier.

The data collection process for an LOE should be tightly coupled to the LOE scenario design process, since that process identifies specific crew behaviors and actions to be expected in a given event set. Pilot performance on the LOE is the ultimate validation of an AQP curriculum, since the LOE is intended to test a representative sample of both technical and CRM skills on which proficiency should have already been established during preceding curriculum segments. If aggregate pilot performance on LOE event sets shows an unsatisfactory trend, the carrier data collection process should be detailed enough to identify where the weaknesses lie. Carriers should develop an LOE tracking system for each crewmember to prevent the same LOE from being used in consecutive evaluations.

d. Line Evaluations

Line evaluation data are the same as LOE data, but on a generic basis, since there is no scenario design process involved. For supervised operating experience, data collection should include verification of proficiency on all TPOs (i.e., sign-off on each TPO by a line evaluator) for the individual crewmember being evaluated much like a validation session. However, regular line evaluations must integrate technical and CRM evaluation for a *full crew*. Aircraft make/model identifiers, crew position, and PF/PNF information is required. If a multi-segment line check is employed, an identifying code to that effect also is required. Line evaluation data should include measures of approved AQP *currency items* that occur during the particular flight. (It is recognized that all currency items may not occur on every flight.) Data also should include a geographic identifier, especially if a random (unscheduled) line check procedure has been approved for the carrier.

e. First Look (Continuing Qualification Only)

First look data are pilot *proficiency data* collected on a subset of maneuver-oriented events that are observed *prior* to any specific performance training on the events. The events used in first look are selected to be diagnostic of skill degradation attributable to insufficient practice. Proficiency on first look events should be judged relative to the approved TPO/SPO qualification standards. First look is applicable only to continuing qualification pilot training and is intended to provide some indication of whether the interval between training sessions is sufficient to support maintenance of proficiency in rarely exercised safety-critical tasks. In addition, it is used to validate that individuals are maintaining proficiency on those items designated as “currency” items by virtue of their practice as part of their normal duties. Analysis of first look data may indicate whether some type of additional training intervention is needed to maintain proficiency during an AQP evaluation period.

f. Instructor/Evaluator

Instructor/evaluator data can be categorized into two distinct groups: 1) data on the instructor/evaluator as a *student* in an instructor/evaluator qualification or continuing qualification program and 2) data on the instructor/evaluator as a *trainer/rater* in a crewmember AQP curriculum.

Both performance and knowledge data should be collected on instructors and evaluators as students in qualification and continuing qualification curricula, as appropriate. Data collected are similar to crewmember data in a pilot AQP, but should be specific to the qualification standards established for the respective category of instructor/evaluator and curriculum segment involved. The data collected will be used by the carriers to track their instruction and evaluation abilities, while their performance as crewmembers will be included in the pilot proficiency database.

In addition to data collected during instructor/evaluator differences training, or during a full instructor/evaluator AQP curriculum, data from periodic observations by instructor or evaluator evaluators (e.g. Quality Assurance

evaluators) of instructor/evaluator performance is required, appropriate to the curriculum segments in which such observations occur.

When instructors and evaluators are performing their jobs as raters in an AQP session, instructor/evaluator identifiers should be included on all rating forms used on students in training. Periodically, average ratings and standard deviations by item should be compared between different instructor/evaluators for the same curriculum segments as an indication of standardization in instructor grading criteria.

As in Phase III activities, carriers should record evaluator-identifying information on gradesheets for quality control. This will also permit the carrier to conduct their own internal "Inter-Rater Reliability." The FAA inspector's name or identification number should be recorded on gradesheets whenever an evaluation is observed.

4. Data Collection Summary

The table below summarizes the recommended types of data to be collected for the various AQP curriculum activities during phases IV and V. (The data collection requirements during AQP phase III are straightforward and are not included in the table.) For simplicity, ancillary information is not included in the table. However, in all activities, ancillary data should be collected to allow proper identification of the aircraft, curriculum, type of session, instructor/evaluator providing the training or evaluation, date, and crew position. The recommended data types apply equally for both qualification and continuing qualification, except for the training activity, which has separate requirements for qualification and continuing qualification, and the first look activity, which applies to continuing qualification only.

Data in this table are recommended for carrier internal use, but are not required for reporting to the FAA. FAA reporting requirements are listed in Section 13.

Data Type	**Training**		**Validation**	**LOE**	**Line Eval**	**First Look**
	Qual	**Cont Qual**				
Performance Data						
PF/PNF				X	X	
Regional Area					X	
Multi-Segment					X	
Seat Substitute		X	X	X		X
Overall Rating		X	X	X	X	
Ref. Addl Trng		X	X	X		X
Elim from Trng			X	X		
Event ID	X	X	X	X	X	X
Event Rating	X	X	X	X	X	X
Reps to Prof.	X	X	X			
Reason Code		X	X	X	X	X
Enabling Data		Possibly	EO Only			

SECTION 10. DATA ENTRY

1. General

All performance/proficiency data collected throughout the AQP should eventually be entered into the performance/proficiency database. Typically, this is some sort of electronic database that will ease the functions of analysis, comparison, and reporting. Considerations for data entry include the method, the hardware/software required for data input, storage, and manipulation. Distinct advantages, disadvantages, and costs are associated with any method of data entry selected, and each individual carrier should accomplish its own cost benefit analysis to determine

the course of action best suited to its training program. A summary of the various methods for data entry along with the considerations involved in selecting a data entry method is shown at the end of this section.

2. Performance/Proficiency Database

Performance/proficiency data collected in the AQP and SVT programs forms the performance/proficiency database. The contents of the database are continuously maintained by the carrier and added to regularly as new data are received. The data within the database can be grouped in several ways (total carrier data, by fleet, by crew position, by geographic area, etc.) for analysis purposes.

While the aggregate performance/proficiency information is termed "database," there is no FAA requirement to input the data into an electronic table. Carriers have considerable latitude in designing a workable performance/proficiency database. The database may include hard copy records, electronic records, or a combination of both. However, the FAA does require regular electronic reporting of the proficiency data, Consequently, the FAA Manager of AQP *strongly recommends an electronic database using a standard off-the-shelf software product.* This allows carriers to make detailed queries of their performance/proficiency data easily on a regular basis for analysis, trend reports, etc. These analyses and reports primarily are used internally by the carrier for student and training program validation. Carriers may also make higher level queries of their data to provide summary reports to the FAA on a regular basis. The remaining discussions in this Guide assume an electronic performance/proficiency database.

The FAA also maintains an electronic database containing the performance/proficiency data of all carriers involved in AQP and SVT. The data within the FAA's database are supplied by each carrier electronically and should be a subset of the data within the carrier's own performance/proficiency database. Reporting requirements to the FAA are described in a later section of this Guide.

3. Database Hardware/Software

As previously recommended, for data to be manipulated easily, they should be input and stored in some form of electronic database. This requires only a common desktop computer and a database software package.

The desktop computer should be capable of accommodating a reasonably complex database management system, contain a large hard drive for data storage (minimum 1 GB recommended), and incorporate some method of making regular backups. Because of the large amount of data that will accrue over time, removable cartridge, tape, or optical media is recommended for backup. The FAA Manager of AQP has selected and recommends a high-end Pentium processor workstation and Windows 95/Windows NT operating system environment. This configuration should support current and future database requirements of most SVT/AQP participating airlines.

Several off-the-shelf database software products are available, all of which are very adequate at storing and manipulating large amounts of data. Some of the more common types used by participating AQP carriers include Access, FoxPro, Paradox, and proprietary software. Costs for each product depend upon several factors:

- Stand-alone application versus a software part of an integrated office suite
- Single-user versus multiple installations
- Site, network, or unlimited licensed copies
- Developers Toolkit add-ons

Carriers should consider the following factors when deciding which application(s) to use, including:

- The Enhanced Model AQP Database, which is being developed in Microsoft Access 97
- Size and complexity of the carrier and/or the carrier's fleet

- Resources available to the carrier
- Type of automation available to the carrier
- Current standard software applications being used by the carrier

4. Data Entry Methods

Generally, three methods exist for entering data into the performance/proficiency database: fully automated, manual, and semi-automated. Most carriers use some limited combination of these three methods.

a. Fully Automated

The fully automated system is one in which data are collected and input into the database directly via a computer with little to no human intervention during the collection/input process. Automated input has many advantages: scoring is completely objective and uniform for all people being trained or tested, the same question sequence and scoring rules are applied uniformly to everyone using the device, data are consistent with respect to the user's responses, the process is completely deterministic, and each pilot who responds identically gets the same score.

Automatic data collection and input is used primarily in ground school to collect EO information via computer-based training. Computer-based training and testing terminals can collect a variety of information automatically, such as lesson completion times, test scores, answers to questions, time required to answer each question, etc. When the session is over, the computer can tally up the score and record the results. The results can be printed on paper for the examinee, to a file for the examiner, or directly to a database. Although the major knowledge-assessment information consists of answers to each question and overall test scores, the auxiliary information such as study time or response time for each question may also prove to have diagnostic value. If research confirms that the timing information has diagnostic value, it should be considered for inclusion in the carrier's information database.

Simulators and some other flight training devices can be programmed to collect performance data automatically on maneuver parameters and may even be used to assign ratings based on a predetermined set of conditions. However, fully automated data collection and input via simulator is still in the investigative stage and is not widely used.

b. Manual

Manual methods require an instructor/evaluator to observe performance or test knowledge of measured items, render a judgment on the observation, and record the judgment. Most instructors will be judging performance during a simulator session and testing for knowledge during an oral exam, then recording the results on paper media. The instructor or separate data entry operator then manually keys the paper record into a computer.

An advantage to the manual method is that the paper notes can be used for the debrief following the training or evaluation session. Primary disadvantages are that the decision to take action is a judgment call and two people with identical responses can be graded differently. In addition, the actual recording of the decision on paper media can sometimes have erratic results. Strictly manual methods are not widely used, but represent a low cost alternative for carriers with small fleets.

c. Semi-Automated

The semi-automated method of data collection/input is a combination of the automated and manual methods and is the most widely used among larger carriers for collecting performance data. This method entails collection of data manually by the instructor/evaluator, then automated input into a computer. Typical input systems are OMR "bubble" scanner sheets, input via the simulator instructor/operator station (IOS), input through data collection programs on a computer, etc.

Recording in the simulator can be automated to the point of programming the IOS to prompt for ratings, reason codes, etc. based on differences from predetermined parameters. While the simulator computer may be recording these differences, the instructor/evaluator, when prompted, must still make a judgment call regarding the outcome of a particular measured item or event.

The semi-automated method has many advantages. While the decision to take action (assign a grade other than satisfactory and/or a repeat) is a judgment call and two people with identical responses can be graded differently, once the "action" is initiated the recording is automated and automatic. The recordings are a consistent rendition of the instructor/evaluator judgment calls. None is lost by failure to enter into the data capture process.

5. Data Entry Systems

Depending on the method selected by the carrier for data entry, specialized entry hardware and/or software may be required. Data manually keyed directly into a database require only the database hardware and software for entry. However, proprietary software may be required for other, more automated, forms of data entry.

a. Instructor/Operator Station (IOS)

Data entry via the IOS console in a simulator or flight training device or data entered via a carrier-wide flight operational computer may require specialized programming by the device vendor or in-house software engineers. While this type of entry may be advantageous in many instances, it is usually very expensive and difficult to modify later. In addition, entering large amounts of data directly into an IOS or flight operations computer can be very cumbersome and time consuming for the instructor/evaluator. Consequently, no carriers currently use automated data entry via the simulator or flight training device computer for AQP/SVT data collection. However, a few carriers use their operational flight computers for entering a limited amount of performance data.

b. Onsite Computer Entry

Data entry via a data entry-supported program on an on-site computer incorporates the advantages of manual data entry and also supports automated data entry error checking.

c. Optical Mark Reader (OMR)

Automated entry of performance data via optical mark reader (OMR) equipment is relatively inexpensive, easy to maintain, and relatively easy to modify. OMR data entry requires specialized hardware (a scanning device), usually specialized software that is unique and proprietary to the device, and OMR forms (bubble sheets). A disadvantage of using the OMR data entry method is that data errors occur more frequently.

d. Flatbed Scanner

A relatively new type of automated data entry system being considered is similar to the OMR method, but somewhat less complicated and requires a less expensive hardware suite and forms production process. Forms are defined using proprietary software, mass-produced via a simple black and white copy machine or printer, and then a flatbed scanner reads completed forms. This method has the potentially added advantage of being able to automate the input of comments as well as rating data.

SECTION 11. DATA REPORTING

1. General

Reporting of data is based on the computer analysis of the performance/proficiency database to provide information on the curriculum and participant groups. The primary users of data reports are internal carrier personnel and the FAA. This section of the Guide makes recommendations for internal airline reporting. A following section describes the external reporting requirements to the FAA.

2. Internal Reporting

The primary use for data collected within the training program is to generate internal carrier reports. The format and frequency of generating internal reports are unique to the needs of each carrier. Ideally, the reports should be fully automated so that fleet managers, instructors, and course designers can obtain relevant statistics on a continuing basis.

Users should not have to design their own data query and report format. However, a data analyst should be readily available to assist the users in interpreting reports, designing new reports to satisfy changing management report requirements, and conducting follow-up analyses when significant results are found. The analyst should be pro-active and meet regularly with the key managers and personnel within the training development and training management systems to review their report and data collection requirements.

3. Report Formats

Generally, the following three basic kinds of report formats are available, each with its own advantages and disadvantages. Most carriers use a combination of the three in various styles.

a. Written

All reports should have some written text. Do not assume that reviewers will look at a pile of data presented in tables, charts, and graphs and reach the same conclusions. Therefore, written text is required as background information, to point out conclusions, assumptions, methods, special conditions, etc. Anything that needs to be brought to the readers' attention should be presented via an accompanying textual report. Any tabular or graphical data presented in the report should be used only to support the written text.

If a particular analysis series is routinely performed on a regular basis, it may be possible to treat each iteration as an addition to a main report that outlines the background, assumptions, methods, etc. Conclusions and any special conditions could be part of each iterative report. Periodically, the main report should be reviewed and updated in its entirety.

b. Tables

Tables are an effective way to present a large amount of data in a compact space and are ideal for spreadsheets and other numerical analysis techniques. The tabular format also permits the display of some intermediate results to help give a sequence of important changes to the individual reviewing the data. Significant values can be highlighted for emphasis.

The disadvantage of tabular presentations is that they are not necessarily intuitive. The reviewers must have sufficient experience to visualize and establish in their own mind the relationships between data points.

c. Graphs & Charts

Graphs and charts are used to present data visually. The relationships between data values are expressed in visual analogues, such as pie charts, bar charts, line charts, etc. Pictorial data make relationships stand out and conclusions intuitively obvious.

Graphs and charts are very effective when the point to be illustrated can be presented in a simple diagram. However, the person constructing the graphic should avoid the tendency to get "cute" and add extra spatial dimensions, such as shadows. These constructs provide only visual clutter and no information.

The disadvantage of graphical presentations is that the charts can quickly become cluttered. When sufficient data are presented for analysis, it can become difficult to pick out the significant elements. One partial solution is to use color to highlight elements that require attention. "Red Flag" results (discussed below) can, for example, be colored red. Graphical analysis is essentially manual interpretation.

4. Reporting Software

Usually, the same software used for storing data in the performance/proficiency database can be used to generate reports. Depending on the complexity of the report, some carriers have found it easier to query the database and then to export the results to another software package for analysis and reporting. Some of the more common types of

reporting software products used by several carriers include Crystal Report Writer or the combination of Access, Excel, and Microsoft Word.

SECTION 12. PROPOSED AIRLINE DATA REPORTING

1. Report Timing

This section includes findings from recent AQP grant-funded research conducted between academia and selected AQP-participating airlines. Monthly summaries of training and evaluation results will be provided to the airline Quality Assurance (QA) department. QA will interpret these reports and disseminate appropriate information to training, CRM department, management, and fleets.

2. Report Content

Routine airline reports should plan for both detailed and summary levels of analysis. Additionally, finding specific results in the routine reports will trigger special reports. Criteria for results that trigger further analysis are called "Red Flags."

3. Detailed Reports

For each fleet, detailed reports will analyze each assessed element in a training or assessment event. The assessed elements would be each maneuver on Maneuver Validation, and each rated item on LOFTs or LOEs. For example, the analysis results reported for 4-point scale judgments for each item in an MV, LOFT, or LOE are as follows:

- Item content
- Count of cases in this sample for this item
- Percent of Missing data (blanks or "Missed observation") for this item
- Minimum and Maximum values observed in this sample for this
- Average score for this item
- Standard Deviation of ratings for this item
- Standard Error of the estimated population average for this item
- Number and percent of "1"s (unsat or unsafe ratings) observed on this item

a. Red Flags for Detailed Reports

Routine reports will have appropriate benchmarks for the analysis of each item. Results that exceed these benchmarks are red flagged for further analysis or interpretation. Red Flag criteria that have been developed for the detailed analysis of 4-point MV evaluations are:

RED FLAG CRITERIA FOR SCALE JUDGMENTS IN A MANUEVER VALIDATION	
Analysis Component For The Item	**Red Flag Criteria**
Count of number of pilots evaluated	Should agree with other reports of number of pilots in fleet and scheduling for MVs—should not be too high nor too low
Missing data	No more than 0-1 % missing data for all items (not including "Missed Observation" ratings)
Min, Max	Do not exceed minimum and maximum possible scores for the scale (e.g. must be in the range 1-4)
Average	Does not exceed possible scale boundaries (e.g. must be in the range 1-4)
Standard Deviation	Is not = 0 or nearly 0
Standard Error of estimating population mean	Is no greater than .10
Number and Percent of "1" ratings	Percent observed is no greater than 5%.

Similar Red Flag criteria will be specified for LOFT and LOE 4-point ratings and the 3-point scale of Fully, Partially, and Not Observed behaviors for the LOE. These criteria will be automatically incorporated into the detailed report wherever possible.

4. Summary Reports

Summary reports combine information across items in the detailed reports. Each summary report uses suitable statistics to answer a focal question in the airline Data Analysis Plan. Examples of Summary Reports are:

✈ **Trend Analysis:** Fleet performance trends are analyzed across months. Fleets with significant declines in performance over months are Red Flagged.

✈ **Performance Profiles:** Performance on all items in an assessment will be compared for significant differences and low performance items are Red Flagged.

✈ **Seat Differences:** Comparison of average performance for different crew positions. Positions with significantly lower performance are Red Flagged.

✈ **In-the-door, Out-the-door Analysis:** Comparable items from the LOFT given at the beginning of an AQP CQ evaluation can be directly compared with appropriate LOE items given at the end of the event.

✈ **Cross-fleet Comparisons:** Comparable evaluation events (e.g. common LOFT) or comparable specific items (e.g. V_1-cut on Maneuver Validation) can be analyzed for fleet differences.

✈ **Instructor Calibration:** Systematic differences and congruency of raters may be analyzed from LOFT, LOE, MV, or line check data.

a. Red Flags for Summary Reports

Precise benchmarks or Red Flags for summary reports will depend on the type of summary analysis. Examples are given below:

RED FLAG BENCHMARKS FOR SUMMARY REPORTS	
Summary Analysis	**Red Flag Criteria**
Trend Analysis of average performance in a fleet across months	Non-chance declines in performance marked by significant negative trends will be Red Flagged for fleet reports. Drill down analysis to specific items.
Performance Profiles for items in a training or evaluation event	Items that are significantly lower than the overall average will be Red Flagged for training reports. Low items tracked through AQP PADB to curriculum changes.
Seat Differences	If significant position differences exist, the crew position with lower performance will be Red Flagged. Drill down analysis to Reason Codes for lower performance.
In-the-Door, Out-the-Door analysis of comparable items at the beginning of training and at the end of training	1. In-the-door items may be compared to an incoming proficiency benchmark. If significantly lower, the item content is Red Flagged for additional training. 2. In vs. Out scores on comparable items may be tested. For items that were not satisfactory during the In-the-door evaluation, the out-the-door scores must be significantly higher than In-the-door scores or the item is Red Flagged for further analysis of why training is not increasing scores on this item. 3. Out-the-door items will be compared to a final proficiency benchmark. If lower than benchmark, Red Flag occurs and possible pilot additional training.
Cross Fleet Comparisons (for comparable items or evaluation events *only*!)	Significant fleet differences in average performance are Red Flagged. Drill down analysis for specific items on which fleet differs.
Instructor or Evaluator Standardization	Systematic differences among evaluators must be non-significant or less than the corporate benchmark, or else this is Red Flagged. Average congruency of raters must be greater than the pre-determined benchmark of .70 - .80 or I/E calibration is Red Flagged. Drill down analysis to find the specific instructors who are non-standard.

Similar Red Flag criteria will be developed for critical airline data analyses.

5. Special Reports

Special reports will be constructed as necessary for two purposes: (1) drill-down analyses to follow up the Red Flag results in the routine reports, and (2) special one-shot analyses of auxiliary data that answer a specific airline training or operational question.

a. Drill-Down Analyses

Supplemental reports will contain follow-up analysis for any results in the detailed or summary reports that raise a red flag. For example, differences in Captain vs. First Officer performance can be followed with analysis of differences for specific items and then the analysis of reason codes and comments for poor performance on these items. The results of the drill-down analysis point to ways to fix the problem found in the routine analysis.

One-shot Analyses. One-shot analyses will be conducted for specific airline auxiliary data that will be collected for specific purposes and typically *not* included in any airline database. For example, a small-group tryout of a Quick Reference Card vs. using memory items for a particular manufacturer's aircraft can be conducted with a small sample of crews using each approach during a LOFT. These data would be analyzed separately from the normal airline databases and reported as a one-shot analysis.

A second example concerns the effects of innovations in training or performance measurement tools such as the LOE rating worksheet. A small-group tryout of changes in training compared with an equivalent sample of pilots using a traditional training approach would entail a limited data collection for two groups of pilot trainees. Post-training results for pilots with new vs. traditional training would be compared statistically. These data would be kept separate from the normal carrier database and analyzed as a one-shot study. Changes of the measurement tool such as changing the evaluation scale from a 4-point to a 5-point or even a 10-point scale could be examined in a one-shot study on I/Es during recurrent IRR training. Again one group of I/Es would use the new form and one group the traditional form. The form that produced consistently better IRR results could, barring other factors such as cost, be judged superior and used as the airline standard.

A third example concerns the effect of pilot background and experience on training and performance. Training is required when pilots transition between fleets, and the past experience of these transition pilots may either help or hinder their performance in the new fleet. For example, past career paths that lack a high level of active flying experience may produce poorer training performance. Again a small group tryout of measuring the immediate past experience of pilots entering training can be correlated with training results in a one-shot analysis.

6. Report Formats

The formats for reporting specific kinds of results should be standardized wherever possible. The formats will have standardized content and emphasize either visual or tabular formats for easy comprehension and use. These formats are currently under development by the academic grant team for selected airlines. Proposed format standards are:

✈ **Distributions of raw data:** Vertical bar charts with relative % as the Y-axis. Distributions should be normalized within fleet due to differences in fleet size.

✈ **Comparison of means:** Vertical bar charts with the *entire range of the scale*(Max to Min) as the Y-axis.

✈ **Variance accounted for:** Pie chart with each segment of the pie representing the relative percentage of variance accounted for by each contributing factor.

✈ **Process graphs:** Box and arrow figure. Each box represents a measure and each arrow is labeled with the strength of the specific link.

✈ **Tables:** Tables can be used to summarize information that is too complex for an effective graphical presentation. An example is a listing of results for all the evaluators in a fleet, which may require 20 or more distinct rows in the table.

Other formats will be developed using appropriate principles of meaningful verbal description and data visualization to convey the meaning of the results in the above reports.

7. Report Process

Results from these analyses must be reported, as appropriate, to managers and fleet personnel. *However, this reporting must be done correctly or the integrity of the entire AQP data collection can be jeopardized!* For example, if the report to a fleet simply emphasizes that the percent of "unsatisfactory" grades is too high for a specific item, the easiest way to "fix" this problem is for the fleet I/Es to change their internal grading criteria so that they no longer give any unsatisfactory grades. This is an example of Systematic Error or bias in the evaluation process and potentially makes AQP results unreliable or even invalid for this item.

Another problem encountered in reporting unfavorable results is a tendency to either "shoot the messenger" or "explain away" the results by a series of *ad hoc* critiques of the results. The tendency to explain away unfavorable results defensively and therefore not to deal with them actively is a particular problem for the long-run effectiveness of the AQP process. In order to avoid or at least to minimize these problems, reporting guidelines are being developed as part of database research sponsored by the AQP office.

A provisional set of guidelines for the reporting process based on current experience is:

✈ **Avoid Reporting Bare Facts Alone:** Bare reporting of unfavorable results can backfire on the AQP data collection process as explained above. Unfavorable results should be put in a context of total fleet performance and presented in person (if possible) so that managers and fleet personnel can interactively ask questions of the data analyst.

✈ **Non-Confrontational Process:** Unfavorable results should not be used to lambaste the fleet managers or trainers about "What did you do wrong?" Favorable results should also be given emphasis in summary reports to avoid this type of confrontational process.

✈ **Problem-Solving Solution Context:** Unfavorable results should be addressed as problems that should be jointly solved by management, fleet personnel, pilot representatives, and other relevant parties. All parties should work together to understand the problem further and to devise appropriate solutions.

✈ **Avoid Defensive Denial:** Affected personnel will come up with a variety of *ad hoc* explanations of the form "That result isn't real—you're just finding that result because of X, Y, or Z." For example, if Captains are performing worse than First Officers on crosswind landings on Maneuver Validation, the *ad hoc* explanation may be that the I/Es intentionally make the landings more difficult for the Captains by, for example, cranking in more crosswind. If reasonable, these explanations must be checked out either by talking to the I/Es, or in some cases by further analyzing the data. *However, the results should be treated as real until such time as the alternative explanation has empirical evidence*. It is far safer to err on the side of doing something when it may turn out not to be necessary rather than make the mistake of not doing something about a real performance problem in a fleet which may result in an accident.

✈ **Have a Clear Line of Responsibility:** A problem that is everybody's problem often winds up being nobody's problem. Have a single person or group designated to investigate the alternative explanations for the unfavorable results and/or to implement solutions developed by the group to see if the suggested solution

solves the performance problem. This person or group should have the responsibility to report back on their findings to all relevant parties at a specific future date.

These guidelines will be elaborated as experience with reporting AQP results is accumulated by research teams working with designated carriers.

SECTION 13. FAA REPORTING

Data Submission

The FAA has established the minimal requirements for submission of de-identified data by curriculum as indicated below. Data should normally be submitted in one calendar month blocks. These submissions should be forwarded electronically to the FAA Manager of AQP in an Open Data Base Connectivity (ODBC) compliant format normally within two months of collection. (For example, June results would be expected by September 1.) It is expected that the data collection conducted by the air carrier for its own use in monitoring curricula will support more analytical detail and diagnostic function than the data collected for submission to the FAA.

File Name
The following file naming convention should be used for data submissions:

XYZmmyy?.ext, where
XYZ is the identifier assigned to the carrier,
mmyy is the month and year for the reported training month,
? is reserved for a report "type" identifier, and
.ext represents the proper extension for the applicable ODBC file.

Example: TWA0997A.mdb, for TWA September 1997 (report type "A") in Access 97 mdb format.

1. SVTP Data Collection

SVTP data submission requirements are contained in the individual carrier's SVTP plan.

Fields: Using a Proficiency Rating Scale and Reason Categorization System, the following generic requirements are suggested for First Look, Proficiency Checks, and Line Checks:

1. Single identifiable electronic record, with identifiers to track individual performance for crew position: PF, PNF (depending on form design) whether first look through proficiency training and evaluation. The form itself may be a data collection tool, but could also be used as a record of training.
2. Data collection period—Month and Year
3. Aircraft make: series and variant if significant operational differences (AC 120-53)
4. Type of evaluation

- First Look**:** first occurrence, rating scale and reason code
- Proficiency training: number of repeats where initial performance below standard
- Proficiency check: rating scale for each repetition of a PC maneuver and for each evaluation below FAA standards (Report each unsatisfactory PC evaluation)
- Line Check: limited to PIC (T, M, P)—Report each unsatisfactory line check

5. Simulator session (as identified in the curriculum)
6. Crew position: PIC, SIC, FE
7. PF/PNF, as applicable, per item
8. Seat position substitution, other than normal seat
9. Overall Rating
10. Individual task, maneuver, procedure (T, M, P)
11. Event (T, M, P) Rating
12. Reason Code
13. Number of repeats required to proficiency for First Look and Proficiency Check items
14. Additional training sessions (beyond scheduled simulator session)
15. Referred to Committee
16. FAA simulator identifier number
17. Evaluator ID #
18. FAA identification number (Form 110A number) of FAA inspector observing session, if any
19. Comments - Optional

Summary of SVTP Data Submission Requirements

Data Type	First Look	PC	Line Check
Record Identifier	X	X	X
MM/YY	X	X	X
Aircraft Fleet ID	X	X	X
Type of Evaluation	X	X	X
Simulator Session	X	X	
Crew Position	X	X	X
PF/PNF	X	X	X
Seat Substitute	X	X	
Overall Rating	X	X	X
Task/Maneuver/Procedure	X	X	X
Event Rating	X	X	X
Reason Code	X	X	X
# Repeats Required	X	X	
Addition Trng Required	X	X	X
Referred to Committee		X	
FAA Simulator ID #	X	X	
Evaluator ID #	X	X	X
FAA Inspector ID #	X	X	X
Comments - Optional	O	O	O

2. AQP Qualification Curriculum

The applicant must report the following minimum AQP data to the FAA Manager of AQP:

a. For each crewmember in training, for each progressive validation module other than LOE, by respective validation module identifier:

- Recommended for additional training: Yes/No
- Referred to committee: Yes/No
- For maneuver validation only, FAA simulator identifier number (assigned by the National Simulator Program Manager)
- Unique identifier of evaluator administering session (if an FAA approved check airmen)
- FAA identification number (Form 110A number) of FAA inspector observing session, if any

b. For each crewmember evaluated in an LOE, a single electronic record for the PIC/SIC/FE (or seat filler) identifier codes and for each pilot evaluated:

- PF/PNF
- By LOE identifier code, satisfactorily completed: Sat or Unsat
- By LOE/event set identifier: Assigned grade
- For each Unsat event set: Summary reason code or skill category
- Recommended for addition training: Yes/No
- Referred to committee: Yes/No

- FAA simulator identifier number (assigned by the National Simulator Program Manager)
- Identifier of evaluator administering session (if an FAA approved check airmen)
- FAA identification number (Form 110A number) of FAA inspector observing session, if any

c. For each crewmember receiving Initial Operating Experience (IOE) and line check:

- Duty position identifier (PIC/SIC/FE)
- PF/PNF
- By graded item identifier for each duty position: Assigned rating
- Recommended for additional OE: Yes/No
- Recommended for addition training: Yes/No
- Identifier of evaluator administering session (if an FAA approved check airmen)
- FAA identification number (Form 110A number) of FAA inspector observing session, if any

3. AQP Continuing Qualification Curriculum

The applicant must report the following minimum AQP data to the FAA Manager of AQP:

a. For each crewmember in training, for First Look:

- By First Look event identifier
- Repeat(s) required: Yes/No
- Identifier of evaluator administering session (if an FAA approved check airmen)
- FAA identification number (Form 110A number) of FAA inspector observing session, if any

b. For each crewmember in training, for maneuver validation:

- By event identifier
- Repeat(s) required: Yes/No
- Recommended for addition training: Yes/No
- Referred to committee: Yes/No
- FAA simulator identifier number (assigned by the National Simulator Program Manager)
- Identifier of evaluator administering session (if an FAA approved check airmen)
- FAA identification number (Form 110A number) of FAA inspector observing session, if any

c. For each crewmember evaluated in an LOE, a single electronic record for the PIC/SIC/FE (or seat filler) identifier codes and for each pilot evaluated:

- PF/PNF
- By LOE identifier code, satisfactorily completed: Sat or Unsat
- By LOE/event set identifier: Assigned grade
- For each Unsat event set: Summary reason code or skill category
- Recommended for addition training/evaluation: Yes/No

- Referred to committee: Yes/No
- FAA simulator identifier number (assigned by the National Simulator Program Manager)
- Identifier of evaluator administering session (if an FAA approved check airmen)
- FAA identification number (Form 110A number) of FAA inspector observing session, if any

d. For each cockpit crewmember present during an Online Evaluation: The performance data that directly mirrors the content of the FAA-approved Online Evaluation form to include identifiers for:

- Aircraft fleet
- Duty position
- PF/PNF
- By graded item identifier for each duty position: Assigned rating
- Currency item flag
- Geographic area (if random line check)
- Identifier of evaluator administering session (if an FAA approved check airmen)
- FAA identification number (Form 110A number) of FAA inspector observing session, if any

Summary of AQP Data Submission Requirements

	Training								
	Qualification Curriculums					**Continuing Qualification Curriculums**			
Data Type	Syst Val	Prcd Val	Man Val	LOE	IOE & Line Check	First Look	Man Val	LOE	On line Eval
File Name/Record Identifier	X	X	X	X	X	X	X	X	X
MM/YY	X	X	X	X	X	X	X	X	X
Aircraft Fleet ID	X	X	X	X	X	X	X	X	X
Curriculum ID	X	X	X	X	X	X	X	X	X
Crew/Duty Position				X	X			X	X
PF/PNF				X	X			X	X
Seat Substitute				X				X	
Satisfactorily Completed (Y/N)	X	X	X	X	X		X	X	X
Item Or Event Set Identifier	X	X	X	X	X	X	X	X	X
Event Rating/Grade				X	X			X	X
Reason Code/Skill Category				X				X	
Repeats Required						X	X		
Additional OE Required (Y/N)					X				
Additional Trng Required (Y/N)	X	X	X	X	X		X	X	X
Referred to Committee (Y/N)	X	X	X	X			X	X	
FAA Simulator ID #			X	X			X	X	
Evaluator ID #	X	X	X	X	X	X	X	X	X
FAA Inspector ID #	X	X	X	X	X	X	X	X	X
Currency Item									X
Geographical Area									X
Comments - optional	O	O	O	O	O	O	O	O	O

Syst. Val. - System Validation
Prcd. Val. - Procedures Validation
Man. Val. - Maneuvers Validation
LOE - Line Operational Evaluation

PIC - Pilot In Command
SIC - Second In Command
FE - Flight Engineer
PF/PNF - Pilot Flying/Pilot Not Flying

For each crewmember of a Qualification or Continuing Qualification curriculum, the AQP office must be able to associate the data records applicable for that crewmember in that curriculum through logical grouping of the records or linkage by a common index number.

4. Data Transmittal to FAA

De-identified AQP and SVT data are transmitted monthly, but not greater than 60 days following the month of collection, to the FAA Manager of AQP. The transmittal file should be in a format commensurate with the database or spreadsheet format you are using. The data format used must allow this office to import the data into Access 97. The data sources Microsoft Access can open, import or link are listed in the following table.

MICROSOFT ACCESS 97 DATA SOURCES	
Data Source	**Version or Format Supported**
Microsoft Access	2.0, 7.0/95, and 8.0/97
Microsoft FoxPro	2.x, and 3.0 (import only)
DBASE	III, III+, IV, and 5
Paradox	3.x, 4.x, and 5.0
Microsoft Excel spreadsheets	3.0, 4.0, 5.0, 7.0/95, and 8.0/97
Lotus 1-2-3 spreadsheets (link is read-only).	.wks, .wk1, .wk3, and .wk4
Delimited text files	Most files with values separated by commas, tabs, or other characters; must be in MS-DOS or Windows ANSI text format
HTML	1.0 (if a list); 2.0, 3.x (if a table or list)
Fixed-width text files	Most files with values arranged so that each field has a certain width; must be in MS-DOS or Windows ANSI text format
SQL tables, and data from programs and databases that support the ODBC protocol	For a list of supported ODBC drivers, search the Microsoft Knowledge Base for article number Q140548, "List of ODBC Drivers Installed by Product."

5. Required Documentation for Format Approval

Each carrier must provide the FAA AQP office a detailed description of its proposed data transmittal format for approval before data are transmitted. This description should include data field name, position, length, description, and definition of legal values. Each discrete record layout must be described. *Data collection forms by themselves do not constitute documentation of the transmittal data.*

6. Transmittal Instructions

The transmittal media can be 3.5" IBM PC compatible 1.44MB diskettes by U.S. Mail or courier, via e-mail as an attached file using the Internet, or through File Transfer Protocol (FTP).

SECTION 14. MEANINGFUL DATA ANALYSIS

1. General

This section of the guide describes properties associated with data analysis in an AQP environment. It also suggests new techniques for analyzing the data, and presents examples used in analyzing the data. These new techniques were developed during a research project sponsored by the FAA Manager of AQP in association with academia and selected AQP carriers cooperating in this project.

Once the performance/proficiency data have been identified, collected, and entered into the data management system, an analysis should be performed on the aggregate information. Statistical analysis of the proficiency data enables individual carriers to make an internal assessment of their performance over time.

It is important to point out that, while this section of the Guide makes suggestions that may be adequate for initial analysis of AQP data, individual carrier's should tailor these processes and techniques to suit their own requirements. Each carriers data collection and performance assessment processes should be refined over time, based on its own empirical experience. That is, the measures and processes should be optimized over time so as to provide the appropriate degree of discrimination in crewmember performance desired by each carrier.

2. A Logical Sequence for Meaningful Data Analysis

Four stages that a typical airline might go through to answer questions are illustrated below. This process depends upon collecting relevant, good quality data and effectively managing the information system.

MEANINGFUL DATA ANALYSIS OVERVIEW	
Form Question	Questions air carriers ask should be relevant to operational concerns or to important training issues. These questions must be tied to appropriate AQP data in order to get answers.
Get Good Data	Quality data involves looking at several parameters: Sensitivity (i.e., whether a 3-point, 5-point, or Sat/Unsat scale is used) Reliability (i.e., Inter-Rater Reliability training) Validity (i.e., cross-checking with different forms of measurement) Good Data Collection Process: Completeness & Standardization
Analyze Data	Using both descriptive analysis (e.g., reports and basic information) and inferential analysis (using statistical measurements to answer questions). The difference is in purpose and focus.
Answer Question	Involves descriptive and inferential analyses. Report process must be appropriate. The user must understand the answer to the question and its implications through succinct verbal descriptions, charts, graphs, and figures.

3. Eight Steps to Meaningful AQP Data Analysis

Step 1: State Initial Question or Problem Precisely.

Reduce a general problem into one or more component questions.

Sample problem: The percent of initial qual LOE failures this month is too high!

Question 1: Is the percent of initial qual LOE failures *really* higher than last month?

Question 2: If it is real, *why* are LOE failures increasing?

Question 3: *How* do we fix it?

Rephrase the initial questions into specific questions that have one specific focal point. Typically this focal point will either describe some aspect of the data (descriptive) or make some inference based on the data (inferential).

Rephrasing General to Specific Questions		
Question	**General Question**	**Specific Question**
1	Is the percent of initial qual LOE failures *really* higher than last month?	Is this months percent a non-chance increase over last month?
2	If it is real, *why* are LOE failures increasing?	*On which specific items* are the pilots having lower performance?
3	*How* do we fix it?	*What* specific training interventions would increase performance on these items?

Examples of descriptive and inferential questions are listed below.

Descriptive Questions	
How are the data distributed?	What is percent of failures in First Look Maneuvers Validation? What is the average continuing qualification LOE score?
Profiles	What is the performance profile across First Look Maneuvers?
Patterns	What are the underlying dimensions of LOE performance? (e.g., are there separate dimensions of CRM and Technical performance? What pattern of training results would identify a “cowboy” pilot?

Inferential Questions	
Differences between groups or measures	Are Captains different from F/Os on Continuing Qual assessments? Are Fleets different in average First Look maneuvers performance?
Differences between items or measures	Which Observable Behaviors on the LOE are significantly higher or lower than average? Is the average Line Check grade higher or lower than the average Maneuver Validation grade?
Relationships among measures	Does MV performance correlate with LOE performance?
Trends over time	Is Line Check performance increasing or decreasing over months?

Step 2: Find Relevant AQP Data for Each Specific Question.

Use the AQP databases: PPDB for pilot assessment, PADB for pilot training, and Instructors/Evaluators database (IEDB). Are all required data present in *one* database? Examples of required information are:

- ✈ Differences: Can necessary groups be identified?
- ✈ Relationships: Are the correct measures present?
- ✈ Trends: Are data tagged with times like month and year?

If yes, go on to *Step 3.* If no, are all required data present, but in *different* databases?

- ✈ If yes, appropriately connect the databases and go on to *Step 3.*
- ✈ If no, this question cannot be answered with the AQP data currently being collected—consider adding the necessary information to the appropriate AQP database.

Step 3: Check the Type of Data and How Each Item was Scored.

Types of Data	
Dichotomies	Sat vs. Unsat, Pass vs. Fail
Categories	Reason Codes
Ranks	Ordinal scales
Multi-point interval scales	3, 4, or 5-point scales

How is each item scored (e.g. minimum possible value, maximum possible value, or missing data value)? If this is a multi-point scale, what is the meaning of high vs. low scale scores?

- ✈ If 1=repeat, 2=debrief, 3=standard, 4=above standard, then low scores are poor performance and high scores are good performance
- ✈ For repeats required, low scores (0) are good performance and high scores (1+) are poor performance

Step 4: Assess the Quality of the AQP Data.

Garbage In = Garbage Out, so you must ensure you are working with sensitive, reliable and valid data.

- ✈ Sensitive: Would small changes in performance show up in different scale ratings?
  - ➢ Sat/Unsat ratings would have low sensitivity
  - ➢ Medium sensitivity would be 3, 4, or 5 point scales
  - ➢ High sensitivity would be 6 or 7-point scale
- ✈ Reliable: Low error of measurement, consistent stable results
  - ➢ Inter-Rater Reliability: How good or bad? Exceed benchmarks?
  - ➢ Internal Consistency of a set of items measuring the *same thing.* Want coefficient alpha to be >.70 to .80

- Validity
  - Internal (Structural) validity: Do components or items on this measure inter-relate with each other in the expected manner? For example, on an LOE, do Observable Behaviors correlate with Technical and CRM judgments, and do technical and CRM judgments correlate with overall ratings for the event set?
  - Criterion validity: Does this measure correlate with accepted criteria? For example, do Maneuver Validation grades correlate with Technical Performance grades on the LOE?
  - Construct validity: Does this measure have a reasonable, expected pattern of relationship with other AQP measures? Do training results predict final AQP evaluations?

Step 5: Create a Data Table for the Question.

Use the databases to create a data table that has "The truth, the whole truth, and nothing but the truth." Include all *relevant* information but omit unnecessary information. Join necessary database tables to get ALL the information. Select the required fields to get the *relevant* information. Sort or filter the cases as necessary to get the groups you want.

Check the table description of the item. Does the item description match what you wanted? If you have any doubts, check the raw data collection form to be sure. Check the number of cases. Is it about what you would expect? For example, 1200 pilots in an AQP fleet should give you about 100 recurrent LOEs per month. Too *many* or too *few* points to problems and must be resolved. Perform a preliminary scan by looking over a few rows of the table to check content of fields. Do the fields have the right numbers or letter codes for each item? Resolve any minimum or maximum values exceeded. If none exist, export the table to a spreadsheet like Excel for cleaning.

Step 6: Clean Data (using a spreadsheet like Excel).

Check the extent of missing or out-of-bounds data (greater than maximum value or less than minimum value). If missing data are greater than a minimal amount, check for patterns of missing data across items or cases. For example, extensive missing data may indicate an item was not applicable, or case records for Fill-In pilots in a MV could be all missing data. Delete cases with missing data or estimate missing data.

Check distribution of item values with the *desired* or *expected* distribution of values (e.g. for a 4-point scale, are the % of 1s, 2s, 3s, and 4s about right?). Different types of evaluations may have different distributions even if the same scale is used. Line Checks may have more 3s and fewer 2s or 4s than LOEs.

If necessary, combine or average data. Combine data where appropriate (e.g. create a "Non-Precision Approach" item that is the score of the VOR, NDB, or localizer approach item that was administered by the evaluator).

Many questions require summing or averaging.

- Summing across items for each crew member
  - Sum of "unsatisfactory" levels of performance (1s & 2s) for each pilot
- Averaging across items for each crew member
  - Average for "First Look" maneuvers on the maneuver validation
  - Obtain an average LOE grade by averaging the PIC or SIC grades across event sets
  - Average the Line Check items for each crewmember
- Averaging across *crew members*for certain items:
  - If you want to analyze *crew* performance rather than individual pilot performance, calculate the average score of the members of each crew.

For averaged scores, check reliability, since it is a precision of measurement. For averages over a set of items, you can check the internal consistency reliability of that set of items:

- Coefficient alpha index of internal consistency should be >.70 - .80
- If reliability is too low, check for "bad" items to delete and possible other relevant items to add in.
  - Deleting bad items will raise reliability
  - Adding relevant items of the *same type* can increase reliability

    Don't add apples and oranges!

    Added items should *correlate with* the initial set of items (e.g. adding Maneuver Validation Fixed maneuver items to the First Look maneuver items would add items of the same type that should correlate)
- If reliability is still below benchmark, your measurement is "fuzzy" which causes low statistical power.
  - You are less likely to be able to answer your question clearly.
  - In the future, consider IRR training for this set of items.

Check the distribution of average scores by identifying high and low "outliers." Calculate mean and standard deviation for the item or scale. Calculate Z-scores = (X - mean)/S.D. Then locate cases with Z-scores greater than +3.0 (high outliers) or less than -3.0 (low outliers). (Hint: sort cases on the Z-score.) If you are doing *descriptive* analyses, the outliers are important information. Leave the outliers in the distribution. If you are doing *inferential* analyses, the outliers can throw off the inferential statistics. Before analyzing, you should either:

- Delete outlier cases from the analysis
- Replace the extreme values with the average for that item, OR
- "Trim" extreme values:
  - If too high, substitute a raw score that equals the highest acceptable value.
  - If too low, substitute the lowest acceptable value.

Step 7: Analyze Data.

1) ***Descriptive***. Match data analysis to the specific question and type of data by using database analyses (e.g. Total Access Statistics), spreadsheet analyses (e.g. Excel), or statistical packages (e.g. SPSS, SYSTAT). Basic description of data involves describing the distribution that will depend on the type of data.

- Dichotomous (e.g. What is % pass vs. fail?)
- Categories (e.g. What is relative % of the reason codes for failures?)
- Scales (e.g. What is the average LOE score? How much variability is there in Line Check performance?

The description should be in a form that best communicates to the user: tables, charts, graphs, etc. Monthly, quarterly, or yearly reports may use this descriptive information. Descriptive data will often lead to further "why, which, how..." questions that may require inferential analyses.

2) ***Inferential***. Inferential testing of questions includes checking the following items.

Check Conditions for the Analysis (e.g. Delete Outliers) Then DO It!	
Differences between groups	Dichotomous differences (e.g. Are Captains performing better or worse than First Officers on the Maneuvers Validation?)

	Multiple groups (e.g. Is First Look performance different across fleets?)
Differences across measures	Is average Maneuver Validation performance higher or lower than LOE performance?
Differences across items of the same type within a measure	Which LOE Observable Behaviors are significantly higher or lower? Which LOE Technical skill items are significantly higher or lower? Which LOE CRM skill items are significantly higher or lower?
Relationships among measures	How does LOE performance correlate with Maneuver Validation? How do the Observable Behaviors predict Tech/CRM performance?
Trends over time	Is Line Check performance increasing or decreasing over time?

Step 8: Interpret Answer to Question.

Descriptive questions show the distribution by table or preferably by graph. Inferential questions tell the information consumer:

- ✈ Could this test result be occurring just by chance?
- ✈ If this result is NOT likely to be due to chance at a specific level, then report:
  - ➢ Exactly how likely it is to occur by chance ("significance level")?
  - ➢ Size of the effect
- ➢ Practical importance of an effect of this size

If possible, integrate the descriptive and inferential information into a coherent *story* about the data. Telling a complete story may require asking other questions and performing other analyses (i.e., data analysis cycle).

4. Sample Size

The size of a sample directly influences the stability of the population estimates for descriptive statistics and the precision of statistical answers for inferential statistics. This effect is not, however, linear—that is, twice as large a sample does not give twice as good an estimate or answer. Generally, increasing sample size helps in proportion to the *square root* of the sample size. That is, to get twice as good an estimate, you will need to have four times as large a sample. Therefore, although a larger sample size can compensate for some flaws in data collection, there is a limit and no amount of data can compensate for really poor quality measurement.

Any estimate can be made more reliable by increasing the sample size. This has the effect of reducing the sample variance and causing the sample estimate to become closer to the true population value. Some statistical methods require a minimum effective sample size. Successively smaller samples tend to become increasingly erratic and therefore less reliable.

For inferential statistics, increasing sample size increases *statistical power*, which is the likelihood of detecting effects that are present in the data. Having sufficient statistical power allows the analyst to avoid false positive conclusions or false negative conclusions from the data, and therefore gives a better answer for the original question. For regional carriers with small fleets, statistical power may be a problem and aggregating results over monthly samples may be necessary. For major carriers, statistical power is less likely to be a problem. In fact, certain large fleets may have a great deal of data that give an enormous amount of statistical power. In such cases, the inferential statistics can be joined with measures of practical importance, such as the amount of variance accounted for by an effect, to yield a usable answer for the question.

5. Summary

Good data analysis requires several steps. Any "upstream" errors in this sequence will give poor "downstream results." Each step has clear standards and procedures to maximize data integrity and utility for answering important AQP questions. Air carrier managers and trainers can use "What if" analysis to follow up these analyses and gain

even more information about critical issues. Using sensitive, reliable, and valid data, AQP data analysis will answer critical carrier questions.

- How effective is our training? Where do we need to increase emphasis? Where could we decrease emphasis?
- Are there specific groups of pilots or fleets that are having specific problems?
- Is performance getting better or worse over time? Did our training or safety interventions have the desired impact on performance?

Appendices

A. Terms and Definitions for Items Found in the Data Management Guide

B. Data Entry Summary
A comparison of various approaches to data collection

C. Carrier Summary

D. Sample Data Collection Forms

E. Sample Database Structures

Appendix A

Terms and Definitions for Items Found in the Data Management Guide

AQP	Advanced Qualification Program: A systematic methodology for developing the content of training programs for air carrier crewmembers and dispatchers.
AC 120-54A	The Advisory Circular that details the various aspects of AQP.
CBT	Computer Based Training: Classroom instruction that is performed individually by trainees at computer stations.
CQ, CQT, or CQP	Continuing Qualification (Training/Program): Training that follows the initial qualification on a regular basis.
CRM	Crew Resource Management
Descriptive (data)	Data that are used to describe specific aspect of the records being considered.
EO	Enabling Objective: An instructional objective created at the level of an element, skill, knowledge, or attitude. Example: Describe the functions of the hydraulic system.
Event	A training or evaluation situation comprised of a task or subtask to be performed by the crew under a specified set of conditions.
Event Set	A relatively independent segment of a line-oriented scenario made up of several events.
FAA	Federal Aviation Administration
FAA's Manager of AQP	Reference to the Manager of the FAA National AQP Office (AFS 230).
FE or F/E	Flight Engineer
"Fill-in" or "Seat Filler" pilots	A qualified crewmember who substitutes for a student who is unable to attend an evaluation session, thus allowing the rest of that student's crew to complete their evaluation with a full crew complement.
FL or First Look	"First Look": A specific set of maneuvers that are evaluated in the simulator without prior briefing and which are intended for study to evaluate how the pilot group would handle the problem in an everyday line situation, were it to occur.
FOQA	Flight Operations Quality Assurance: A program that receives and analyzes input of flight operations and noted problems in order to identify and to reduce future occurrences of those problems.
FTD	Flight Training Device: A full scale replica of an airplane cockpit that may not have the motion or visual systems associated with flight simulators.
FTP	File Transfer Protocol: An internet format that facilitates file transfer.
HTML	HyperText Markup Language: A computer language that allows formatting to remain with text and documents.
I/E	Instructor/Evaluator: Instructors and Evaluators in the AQP process, some of whom may be both Instructors and Evaluators.
Inferential (data)	Data that are used to infer specific answers based on the data observed.
IOS	Instructor/Operator Station: A part of a simulator that may permit direct

	input of performance evaluation results.
IRR	Inter-Rater Reliability: A program that is conducted periodically to "calibrate" instructors and evaluators, so that they will rate performance as closely to the same standard as possible.
ISD	Instructional Systems Design
LOE	Line Operational Evaluation: An evaluation of individual and crew performance in a flight simulator conducted during real-time line oriented simulation under an approved AQP.
LOFT	Line Oriented Flight Training: A training session in a flight training device or simulator conducted as a line operation including all phases of flight and not interrupted by the instructor, unless negative learning begins to occur.
MV	Maneuver Validation: A simulator session in which specific maneuvers are performed and evaluated to proficiency.
NTSB	National Transportation Safety Board
OMR	Optical Mark Reader: Sometimes called "bubble sheet" paper, a type of form that can be "read" by a scanner or dedicated machine.
"Outlier"	A term used to identify notable data points that deviate a significant amount from the norm.
PADB	Program Audit Database: A database that is used to analyze the elements of a training program and the supporting task analysis that must be accomplished during any training cycle. It may develop lesson plans and be used to address deficiencies found in performance and proficiency by the PPDB (performance/proficiency database).
PF	Pilot Flying
PIC	Pilot in Command (Captain)
PNF	Pilot Not Flying
PPDB	Performance/Proficiency Database: A database that collects results of performance and proficiency evaluations, and which is used to analyze training programs, to spot developing trends, and to correct any problems that may be noted.
QA	Quality Assurance: A program that gathers information and provides it to managers to improve quality of flight operations.
SFAR 58	The controlling document for AQP.
SIC	Second in Command (First Officer/Co-Pilot)
Sim	Airplane Simulator: A full sized replica of a specific type of airplane cockpit, including both visual and motion systems.
SPO	Supporting Proficiency Objective: An instructional objective created at the level of a subtask. Example: Perform Engine-Out Precision Approach Preparation Procedures.
SPOT	Special Purpose Orientation Training: A training session in a flight training device or simulator designed to address specific training objectives. It may consist of full or partial flight segments and may be interrupted by the instructor.
SVE	Single Visit Exemption: A program that is often a precursor to AQP, which permits a single annual visit for flight training and qualification.
SVTP	Single Visit Training Program:
T, M, P	Task, Maneuver, Procedure

TPO	Terminal Proficiency Objective: An instructional objective created at the level of a task. Example: Perform Engine-Out Precision Approach.

Data Entry Summary					
Data Entry Method	**Cost**	**Ease of Use / Learning Curve**	**Quality of Statistic**	**Applicability to Classroom, CBT, FTD, Sim, Line**	**Required Lead Time**
Manual Input to Terminal I/E uses keyboard and/or mouse to enter data into a computer terminal	➢Relatively small cost to implement: terminal(s) and software development ➢Most carriers have physical and labor infrastructure in place	➢Easy to use, but sometimes time consuming and cumbersome ➢Easy to learn, most people already familiar with technology	➢Time degradation error may be a problem ➢Data entry errors are possible	➢Difficult to use in conjunction with classroom, IOE, and Line Check	➢Short
Laptop/Palmtop Enter data on a laptop or palmtop computer while observing the session, then download to a storage medium	➢Moderate to high cost to implement ➢Most carriers do not have the hardware in place ➢Technical expertise exists ➢Software development required ➢Ongoing support may be logistically challenging	➢Small, portable, and easy to use ➢May be difficult to see and operate in a dark Sim or cockpit ➢Moderate learning curve	➢Eliminates time degradation error ➢Minimizes second party data entry errors ➢May have data entry/omission errors partially dependent on Sim and cockpit environment ➢May be distracting to I/E	➢Technology can be transferred to all evaluation media but may be cost prohibitive in the classroom environment	➢Moderate to high due to hardware purchase and software development
Touch Screen Possible locations include FTDs, Sims, and aircraft	➢Large capital expenditure for hardware and software ➢Ongoing support may be expensive	➢Very user friendly and intuitive ➢Small learning curve for users, but moderate to substantial for support	➢Eliminates time degradation error ➢Greatest chance for statistically pure data from a collection aspect	➢Could be implemented in most evaluation media but may be cost prohibitive in the classroom environment	➢Very long lead time
Hand Written or Manual Entry I/E or pilot records session on form sheet, sends it to a central location, and it is manually entered	➢Small cost to implement ➢Could have extra labor cost for ongoing operations	➢Easy to learn ➢Ease of use dependent on form design	➢May eliminate time degradation error if completed during evaluation ➢Second party data entry error could be a problem ➢Data not timely ➢Forms could get lost or separated	➢Can be used in all evaluation media	➢Small lead time to develop physical resources and hire data entry personnel
"FedEx Style" Hand held input device brought to the session by the I/E	➢Moderate to high cost to implement ➢Most carriers do not have the hardware in place ➢Ongoing support may be logistically challenging	➢Small, portable, and easy to use ➢May be difficult to see and operate in a dark Sim or cockpit ➢Moderate learning curve	➢Eliminates time degradation error ➢Minimizes second party data entry errors ➢May have data entry/omission errors dependent on Sim/cockpit environment ➢May be distracting to I/E	➢Technology can be transferred to all evaluation media but may be cost prohibitive in the classroom environment	➢Moderate to high due to hardware purchase and software development
Data Entry Method	**Cost**	**Ease of Use / Learning Curve**	**Quality of Statistic**	**Applicability to Classroom, CBT, FTD, Sim, Line**	**Required Lead Time**
Simulator Performance data are routed automatically to database or I/E enters grade when Sim stopped to repeat an item. In LOE, I/E may be prompted for grade at end of event sets	➢High capital expenditure ➢Some carriers have hardware in place	➢Moderate learning curve involved ➢User friendly	➢Very accurate ➢Removes human error in recording evaluation ➢Can only record exceedance data ➢Removes I/E judgment ➢Cannot capture human factors data	➢Can only be used in the FTD/Sim environment	➢Substantial lead time involved for hardware and software installation ➢Any software changes may be difficult and time consuming
CBT	➢Low capital expense if	➢Moderate learning curve	➢Eliminates time degradation error	➢Only applicable to classroom or	➢Small lead time if already conducting

Performance data are automatically captured and databased within the CBT environment	committed to CBT; otherwise, moderate to high capital expenditure	➢Simple and intuitive to use	➢Very accurate ➢Removes human error in recording evaluation ➢Collects good statistical data ➢Easy to tie in electronically ➢Collection process designed into the system from start	workstation setting	CBT
Bubble Sheet Scan sheet taken into the session or check by the I/E, then later scanned directly into the system	➢Small to moderate depending on number of data input locations	➢Easy to master and use ➢May be difficult to see in dark Sim or cockpit ➢Considered the industry standard	➢Minimizes time degradation error ➢Provides paper backup if needed ➢Requires second step to input and verify data ➢Forms may be lost or separated if not immediately scanned ➢Errors possible with two step data entry ➢Possible tendency of I/E to grade "standard" if too many parameters are requested/required.	➢Could be used in all training or checking environments	➢Short to medium lead time required

Appendix C

Carrier Summary

Appendix C contains a table summarizing the current status of data management activities among the various organizations involved in AQP. Note that only those organizations actively participating in the Data Management Focus Group and who were able to provide information are represented. All members shown in this table have expressed a willingness to provide support with regard to AQP/SVT data management. For more detailed information on any specific organization's activities, contact the appropriate coordinator listed in the table.

Data Management Focus Group
Points of Contact
May 1998

AIRLINE	DATA POINT-OF-CONTACT	OFFICE TELEPHONE #	OFFICE FAX #	E-MAIL ADDRESS
Airbus	Patrick Alizon	(33) 5-61-93-47-64	(33) 5-62-11-07-40	76214.3453@compuserve.com
Air Transport International	Charles "Jeff" Nauman	(937) 264-4570/4571	(937) 264-0947	ceejayen@worldnet.att.net
Air Wisconsin	Scott Orozco	(920) 749-4120	(920) 739-1325	sorozco420@aol.com
Alaska	Mandy Blackmon	(206) 248-7691	(206) 431-3594	mandy.blackmon@alaskaair.com
Aloha	Tina Marshman	(808) 837-6871	(808) 837-6878	aqfltops@pixi.com
American	O.J. Treadway/Kevin Davis	(817) 967-5193/5484	(817) 967-5127	ojtreadway@aol.com/kdavis_amr@aol.com
American Trans Air	Amy Allen	(317) 390-7761	(317) 243-4897	aallen@iquest.net
Atlantic Coast Airline	Doug Siegler	(703) 925-6110	(703) 925-6290	dougcs@aol.com
Atlantic Southeast Airline	(Paul Jackson)	(770) 844-4430		asatraining@mindspring.com
Boeing	vacant	(206) 662-7720	(206) 662-7812	
CCAir	Tom Leahy	(704) 359-5323	(704) 359-0351	j31driver@aol.com
Comair	Cary Ryan/Anjali Markey	(606) 767-2217/2923	(606) 767-2334/2150	74322.156@compuserve.com/amarkey@fuse.net
Continental	Nannette Pikkarainen/Debbie Obaugh	(281) 553-8220/8226	(281) 553-8229	75702.3700@compuserve.com/dobaug@coair.com
Delta	Matt Humlie/Pat Walsh	(404) 715-1110/1001	(404) 773-1179	matt.humlie@delta-air.com/pat.walsh@delta-air.com
DHL	John "Jack" Vyhnalek/Ed Rutherford	(606) 578-8301 x251/x252	(606) 578-8318	jvyhnale@us.dhl.com/erutherf@us.dhl.com
Executive Jet	Chuck Davis	(614) 239-5521	(614) 239-5589	105224.1261@compuserve.com
Fedex	Danny Korn/Deborah Turpen	(901) 360-4945/4255 (alt # for Deborah Turpen (901) 797-6456))	(901) 360-4955	dlkorn@fedex.com (use daniel.korn@worldnet.att.net to send e-mail with attached files)
Flight Safety International	Ken Kelly	(888) 775-4900/(904) 226-4900	(904) 226-4910	kellyk@tsd.flightsafety.com
McDonnell Douglas	vacant	( )	( )	
NATCO	Stephanie Day	(612) 604-9287	(612)	sdd924@aol.com
Northwest	Kevin Sliwinski	(612) 726-2075	(612) 726-8793	kevin.sliwinski@nwa.com
Piedmont	Rei Torres/Pamela Jefferson	(410) 742-2996 x6524	(410) 742-4071	pdoug@usairways.com (personal 73342.11@compuserve.com)
PSA	Richard Mandras	(937) 454-8179/1116	(937) 454-5897	pembo@worldnet.att.net
Trans States Air	Mark Hoffman	(314) 895-3743	(314) 895-6227	transst2@inlink.com
Trans World Airlines	Carl Halford	(314) 895-5566	(314) 895-6679	cdhalfo@twa.com or richfor@anet-stl.com
United (AQP)	Joy Lanzano/Julie Applebury	(303) 780-5243/5867	(303) 780-5234/5860	lanzano@ualfltctr.com/applebury@ualfltctr.com
(OARS)	Harrison Walker/Mary Forsberg	(303) 780-5281/5881	(303) 780-5234	walker@ualfltctr.com/forsberg@ualfltctr.com
United Parcel Service	Greg Gaddis	(502) 359-8481	(502) 359-8858	flt1gxg@air.ups.com
USAirways	Bob Stafford	(412) 747-1659	(412) 747-1648	stafford@usairways.com
	K.D. Van Drie	(412) 749-3235	(412) 749-9453	kdvandrie@compuserve.com
Data Focus Group Chair	Bob Odenweller	(908) 766-5460	(908) 766-5460	73644.775@compuserve.com
FAA AFS-230	Paul D. Johnson	(703) 661-0277	(703) 661-0274	paul.d.johnson@faa.dot.gov

Carrier Data Submission Specifics

May 1998

AIRLINE	CORPORATE DBMS S/W	FUTURE DBMS S/W PLANS	DATA SUBMISSION MEDIA
Airbus	Access 2.0/97	Access 97	n/a
Air Transport International	Access 97		n/a
Air Wisconsin	Access 97		n/a
Alaska	Access 97 (Informix for AQP/FTMS)	Access 97	e-mail
Aloha	Access 2.0 (for PPDB) Battelle Model AQP (for PADB)	Enhanced AQP DB (in Access 97)	3.5" diskette
American	Paradox 5	Paradox 5 (SVE)/Access 97(AQP)	3.5" diskette/e-mail
American Trans Air	Access 2.0	Access 7.0	3.5" diskette
Atlantic Coast Airline	Access 2.0	Access 97	3.5" diskette
Boeing			n/a
CCAir	Paradox 5 (also has Access 2.0)	Paradox 5	n/a
Comair	JAM (proprietary s/w)		3.5" diskette
Continental	none	Access 97	n/a
Delta	Access 7.0		e-mail
DHL	Access 97		e-mail
Executive Jet	Battelle-based Model AQP Database	Windows NT-based s/w	e-mail (future plan)
Fedex	FoxPro 2.6 (SVT). Visual FoxPro 5.0 (AQP)		3.5" diskette
Flight Safety Int'l	FSI Trng Mgmt System (Excel-based)	Access 97	n/a
McDonnell Douglas			n/a
NATCO	n/a	n/a	n/a
Northwest	Access 2.0	Access 97	e-mail
Piedmont	DEC ADABAS/Natural (proprietary s/w)		3.5" diskette/e-mail
Trans States Air	Access 7.0	Access 97	n/a
Trans World Airlines	Access 2.0	Access 7.0	3.5" diskette
United	FoxPro 2.6	Access 97	3.5" diskette/e-mail
United Parcel Service	Access 97		e-mail
USAirways	FoxPro 3.0	Access 97-Visual FoxPro	n/a
FAA AFS-230	Access 97		n/a

Appendix D

Sample Data Collection Forms

The following forms are representative samples of data collection instruments currently in use at some of the carriers involved in single-visit training and/or AQP. These samples are not all-inclusive. Therefore, not all instruments or organizations are represented. Since the methods of data collection vary widely among carriers, contact the Data Management Coordinator listed in the Carrier Summary for more detailed information on any specific organization.

It is also imperative to note that these samples are intended as guides only, and that data collection instruments unique to one organization may not be applicable in other situations. *The instruments for data collection at each organization must comply with the requirements of your particular single visit exemption and be approved for use in advance by the FAA.*

American Airlines (All SVT, forms as distributed in 1996)
- Line Check/Loft
- Recurrent Training/Proficiency Check

American Trans Air
- Single Visit (Three forms)

Piedmont Airlines
- Maneuver Validation
- LOFT
- LOE
- SOE (second officer)
- On-Line Evaluation

TWA
- First Look/Maneuvers Validation (pilots & F/E)
- Proficiency Evaluation (pilots & F/E)
- LOFT (pilots & F/E)
- Line Evaluation

Appendix E

Sample Database Structures

This appendix includes examples of database structures for two of the carriers represented in the forms appendix. The examples are included more to show two approaches to the documentation needed by the FAA to be able to understand and to analyze the data forwarded on a monthly basis rather than to give a definitive version of a database structure. Needs of different carriers will vary, and the requirements of the Single Visit exemption are likely to require an approach that will be somewhat different from these.

American Trans Air

Piedmont Airlines