NASA/TM–20210000045

Methods for Evaluating the Effectiveness of Programs to Train Pilot Monitoring

Dorrit Billman
NASA Ames Research Center

Randall J. Mumaw
San Jose State University Foundation

Michael S. Feary
NASA Ames Research Center

December 2020

NASA STI Program…in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include creating custom thesauri, building customized databases, and organizing and publishing research results.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question via to help@sti.nasa.gov

- Phone the NASA STI Help Desk at (757) 864-9658

- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM—20210000045

Methods for Evaluating the Effectiveness of Programs to Train Pilot Monitoring

Dorrit Billman
NASA Ames Research Center

Randall J. Mumaw
San Jose State University Foundation

Michael S. Feary
NASA Ames Research Center

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California

December 2020

Acknowledgements

This research was supported by funding from the interagency agreement with the FAA titled “Methods for Evaluating the Effectiveness of Programs to Train Pilot Monitoring” DTFAWA-11-A-80004. The authors appreciate comments on this work by Jolene Feldman, Robert Kotetsky, Vin Parker, and Terence Tyson.

Available from:

NASA STI Program
STI Support Services
Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199

This report is also available in electronic form at http://www.sti.nasa.gov
or http://ntrs.nasa.gov/

Table of Contents

List of Figures and Tables

Acronyms and Definitions

ACRM	Advanced Crew Resource Management
ASAP	Aviation Safety Action Program
ASRS	Aviation Safety Reporting System (NASA)
ATC	air traffic control
CAST	Commercial Aviation Safety Team
CBT	computer-based training
CRM	crew resource management
EO	Enabling Objective
FAA	Federal Aviation Administration
FLCH	flight level change (autopilot mode)
FMS	flight management system
FOQA	Flight Operations Quality Assurance
ICAO	International Civil Aviation Organization
IDA	intentional, discriminative, action
JTA	Job Task Analysis
LOSA	Line-Oriented Safety Assessment
MCP	mode control panel
MEL	minimum equipment list
NAS	National Airspace System
NASA	National Aeronautics and Space Administration
NASA	National Aviation and Space Administration
NOTECHS	non-technical skills
PF	pilot flying
PFD	primary flight display
PM	pilot monitoring
SME	subject matter expert
SPO	Subtask Proficiency Objective
STAR	Standard Terminal Arrival Route
T/D	top of descent
TEM	threat and error management
TPO	Task Proficiency Objectives
VNAV	vertical navigation (autopilot mode)
VNAV PTH	vertical navigation path (autopilot mode)
VNAV SPD	vertical navigation speed (autopilot mode)

Methods for Evaluating the Effectiveness of Programs to Train Pilot Monitoring

Dorrit Billman, Randall J. Mumaw, and Michael S. Feary

This report provides a compendium of methods for evaluation of training programs. It addresses training programs for developing pilot monitoring skills and has a focus on monitoring flight path management. It presents the Sensemaking Model of Monitoring as a framework for organizing the targets of training, and thus also the types of change in performance useful for assessing program effectiveness. It provides guidance for designing program evaluations, which can be tailored to address the specific monitoring topics or content for an application to fleet or airline needs.

Report Summary

Section 1: Purpose of this Report

Improved training of monitoring has been identified as an important route to improved safety and reduction of risk from loss of control (CAST, 2014; FSF, 2014). Training programs aiming to improve skills and knowledge for monitoring will need to be evaluated. The evaluation of the skill of an individual pilot is part of program evaluation and thus many of the measures and methods we propose are informative at the individual pilot level. However, a broader set of considerations is needed for assessment at the program level. This report provides a framework for assessment of programs that train pilot monitoring.

The need for improved training of pilot monitoring was motivated by inadequate pilot monitoring of aircraft state and flight path, which contributed to loss of control incidents and accidents (e.g., CAST, 2014; for a review see Mumaw, Billman, & Feary, 2019a). Our framework for program evaluation is targeted at training for monitoring aircraft state and flight path; however, it is applicable for training programs targeting pilot monitoring in general.

Effective monitoring concerns what the mind is doing, not just where the eyes are pointing. Far more than “where you point your eyes,” the expressions “flying ahead of the plane” and “good airmanship” sometimes refer to this broader meaning of monitoring. Our framework is based on a model of monitoring that identifies the cognitive skills and knowledge needed for effective monitoring performance. These skills and knowledge specify the possible targets of training. We describe how the effectiveness of training programs targeting these skills and knowledge can be assessed.

Section 2: Identifying the Target of Training: Monitoring Definition and Model

Monitoring as Sensemaking. Monitoring is the complex cognitive process of making sense of the current, dynamic situation. The result of this activity is a situation model, the internal representation of the pilot's understanding. The framework for seeking information and interpreting events is included in the situation model. We provide a qualitative model of the complex task of monitoring, the Sensemaking Model of monitoring (see also Mumaw et al., 2020). This model identifies the component processes and knowledge needed for monitoring; in turn, this provides a framework identifying the targets of training. There are four key processes; the first pulls out initial set up of a situation model and the next three define a cycle of updating and deepening the situation model:

1. Initialization sets up the situation model by drawing on the pilot's mental models and other domain knowledge that is recalled from memory; then, a cycle of three processes update the situation model:
2. Identifying the monitoring question that most needs to be addressed.
3. Gathering and assessing the relevant evidence to answer the question.
4. Identifying what, if any, actions are needed in the current or upcoming situation.

Initializing the model depends on accessing the relevant mental models from memory that store the pilot's generally applicable knowledge of how things work. The initial model evolves as additional observations are made, compared to expected values, and projected to the future.

Task Management and Communication. Monitoring takes place in the operational context where both task management and communication are critical. Various monitoring activities in combination with other tasks need to be coordinated. In a multi-crew cockpit, monitoring without communication does little good: the Pilot Monitoring and Pilot Flying need to communicate. Recognizing their criticality to monitoring success is certainly important, whether these are considered as part of monitoring or as ancillary skills.

Targets of Training. The Sensemaking Model of monitoring identifies component skills and knowledge that are the target of training monitoring. Assessing effectiveness of a program for training monitoring then depends on assessing the effective acquisition of these skills. We use a modified version of Kirkpatrick's Levels of Training Evaluation to organize our discussion of relevant evaluation methods.

Section 3: Organization of Evaluation Methods: Levels of Evaluation

Kirkpatrick's levels of training evaluation provide a convenient organizational framework. Kirkpatrick has described four levels for evaluating a training program, and we use a modified form of these levels to organize our reporting of evaluation measures for the various monitoring skills. Level 1 assesses learner attitudes about the training program, such as how useful or interesting they thought it was. In contrast, Levels 2-4 use performance measures. A performance measure assesses observable behavior that can be scored as better or worse. Section 4 describes a large set of performance measures, most of which can be used for multiple levels. These levels were distinguished by the type of evaluation question addressed, but each of these levels was also associated with particular types of measurement and contexts of assessment.

Developments in aviation training technology allow more powerful assessment at lower levels than previously possible. New technology and training methods create new distinctions and new overlap across these levels.

For contemporary aviation training the levels can best be characterized a bit differently:

- As in the original formulation, Level 1 asks about trainee attitudes, particularly concerning the merits of the training program itself; it is usually tested with a survey and typically done wherever training was delivered.
- Level 2 traditionally asked about trainee work-relevant knowledge and concepts in a classroom setting, using tasks and measures such as multiple-choice or short-answer tests of memory for facts. For aviation, technologies now provide a wide array of methods that can assess work-relevant component skills and knowledge. A variety of presentation methods can be used including interactions with videos, procedure trainers, and desktop simulators. A wide range of interactive responses relevant to monitoring can be used, such as viewing and critiquing cockpit videos, providing explanations and predictions, or identifying unexpected values. These options can make Level 2 and "classroom" evaluation assess part-task performance where activities are valid parts of the activities used in actual flight.
- Level 3 traditionally asked if performance in actual, normal work improved. As with Level 2, changes in technology allow changes in assessment; performance in simulators provides important proxies for performance in actual flight. For aviation training, splitting Level 3 into Levels 3A for simulators and Level 3B for actual flight provides a more useful division of training and evaluation practices.
- Level 4 asks if training impacts performance on organizational-level indicators. An organization might identify performance measures relevant to safety that the training program is designed to improve. These might be changes at the fleet level in pilot behavior, in aircraft state, or in reporting of safety issues.

Section 4: Performance Measures are Useful for Levels 2, 3A, 3B, and 4

A. Assessment of monitoring skills relies on two broad types of intentional, discriminative, actions (IDA's): verbal and nonverbal. Monitoring is closely related to awareness and to intention and asking about these is typically a very efficient way of getting the target information. If a nonverbal action would only be taken if the pilot had a particular awareness or intent (it is intentional and discriminative), then that can provide a reliable indicator of monitoring as well. See also Billman, Mumaw, & Feary (2019).

B. Several evaluation strategies apply widely across different performance measures:

- Use a variety of types of measures and tasks at different levels. Assessments in simpler and in more complex contexts have complementary benefits. For example, simpler contexts (often lower levels) are usually more diagnostic of specific gaps in knowledge or skill, while more complex contexts (often higher levels) can be more realistic and provide "whole-task" training.
- Avoid floor and ceiling effects, namely, performance that is too low or too high to measure change. To do this:
  - – use item sets with items that vary in difficulty.
  - – assess the target monitoring component in conditions when that component is likely to limit performance, not some other factor.
- Consider carefully the relation between the particular tasks and items used in training and those in assessment.
  - – Varying the similarity can allow assessment of retention, generalization, and transfer.

    - Coordinating the specifics of the assessment plan with the specifics of pilot evaluation in training can provide efficiencies.

- Design of the scenario or specific use case is critical.
    - This controls whether the specific monitoring skill and knowledge is in fact assessed.
    - Carefully chosen combinations of characteristics of context and of task can aid effective sampling across a large set of possible cases.
- Consider both generative (open response) tasks, like conducting a briefing or recalling information, and choice (closed response) tasks, like selecting one of two modes or recognizing information, because they have complementary strengths and each can be helpful.

C. Measuring Program Effectiveness

Content and Knowledge Access

Relevant, detailed, and consistent mental models contribute to an accurate situation model. Content can be measured through a variety of verbal tasks. These can be done based on static stimuli, videos, and operation in simulator or revenue flight. In revenue flights, probing the pilot after the fact in debriefing or even in cruise may be options. Measures include:

- fact recall or recognition
- explaining a situation
- prediction from a situation
- model selection, including startle and surprise

Assessing the Steps in the Monitoring Cycle

Measuring the three parts of the monitoring cycle can diagnose specific weaknesses in the training program. This can provide specific feedback about where training needs revision. Several response measures can address each of the phases:

1. Question selection.
2. Evidence comparison and interpretation.
3. Action assessment.

For each of these, the pilot can produce a verbal report or choose from a set of response alternatives. In addition, situations and environments can be configured to require a nonverbal action, either manipulating displays or taking a control action. For example, for evidence assessment, a test display can be configured that requires an active click or point to reveal information. For assessment of action, the task can be specified so the pilot takes the needed control action.

Eye-tracking may provide a useful supplementary measure in assessing what evidence is being sampled; however, the link between awareness and gaze is inexact. Eye gaze might aid an expert observer judge pilot awareness or understanding. We have no basis for assuming there is a fixed, cross-situation scan pattern that is optimal and should be used as a training standard.

Assessment of the success of the overall cycle rather than a specific phase is somewhat easier, as it depends on all the phases being executed fairly well. As with assessment of component skills, overall assessment can be measured in a variety of contexts, from paper scenarios, through videos, simulation, and actual flight. Because so much of monitoring does not naturally and necessarily produce observable behavior, it can be particularly helpful to use controlled situations where a correct response is known and maps onto a clear behavioral indicator.

Careful design of scenarios (whether on paper or in a simulator) or identification of types of scenario encountered (in revenue flights) is very important for scoring monitoring performance, and particularly so if the goal is to assess specific components of the monitoring cycle. Of course, selection of the scenarios that are most important to assess is critical. Identifying situations that regularly pose monitoring challenges is an important part of developing a particular training evaluation within this framework.

Assessing Task Management and Communication

These skills in particular may benefit from assessments that include higher levels of evaluation, and where events unfold dynamically, and the overall setting is as natural as possible. Nevertheless, activities such as planning task-allocation or role-playing communication can also be helpful.

Sections 5, 6, 7, 8, and 9: Example Evaluation Measures and Methods by Level of Evaluation

A. Level 1: Participant Attitude

Measuring participant attitude toward training can provide a useful prediction about training acceptance. Participant feedback can also be used to identify parts of the training judged unclear, too repetitive, or otherwise problematic.

Surveys are the usual method for collecting attitudes. Surveys are a well-developed method with known design strategies, for example, to minimize response bias and to increase the odds that participants will provide complete and thoughtful responses.

B. Level 2, 3A, 3B, and 4

We provide examples of how performance measures can be applied at each level:

- Level 2 evaluations can be highly diagnostic. Many relatively quick and inexpensive evaluation items can be completed to provide a sharp view of the target monitoring skill or knowledge. Presentation technology and interactive tasks allow foundational assessment of the monitoring components.
- Level 3A, evaluation in simulated work conditions, is particularly useful because of the combination of control and realism it offers. For Level 3B, performance measures are primarily based on observation of pilots, as in Line Oriented Safety Assessments (LOSA) or in airline-specific observations or check rides. Information captured from the simulator or the aircraft (as in Flight Operations Quality Assurance [FOQA]) also provide data sources. Some established observational coding schemes do address topics strongly related to monitoring, such as situation awareness and communication, but development of standardized coding for monitoring skill is an important topic for development. An additional source of data could be pilot surveys or reports, for example, noting challenging situations or pilot-experienced training gaps.
- Level 4 concerns operational impact, in particular, improvement in safety at an organizational level. Although incidents and accidents linked to inadequate monitoring are a key motivation for attempting to improve monitoring training, these are very insensitive measures. Thus, proxy measures for better monitoring or improved safety are needed. There are several sources for relevant data from which measures of operational safety goals could be selected or developed. 1) Data for some operational

goals might be based on measures of pilot performance, such as increased compliance with particular procedures, e.g., for briefing methods or stabilized approaches. For these safety improvement goals, FOQA data could provide useful information. Data from observer score cards could also be used. 2) Measures already in use by the Safety Management System might be close enough to the training goals to merit use in training assessment. 3) New measures could be derived from FOQA or from established observations, or their combination. 4) Safety reporting could be another source of data. For monitoring in particular, there may not be operational measures that directly reflect better awareness; the impact of training greater situational understanding may be diffuse.

Section 10: Does the Training Program *Cause* the Desired Improvement?

This report addresses evaluation of training programs for monitoring. The key question here concerns whether and how much the training produces improvement in performance. Two types of comparisons are relevant to answering this question. One compares the same person's performance before and after training (within-subject), while the other compares performance of pilots who have taken the training to similar individuals who have not. Both have limitations. A combination of these, in a 'mixed design' provides a very powerful methodology. We lay out how to execute these designs and describe the tradeoffs.

Gathering the relevant information requires careful advance planning. Planning, plus the ability to coordinate with existing practices for collecting data (such as First Look), can produce powerful evaluation opportunities with relatively modest incremental cost. There are feasible, efficient methods for reasonably determining whether a training program caused a benefit and for characterizing the nature of benefit found.

Section 11: Method Summary

Dimensions of Evaluation Methods, with Examples
We describe what combinations of Evaluation Context, Materials, and Tasks are likely to be generally of high value. For example, we suggest high-value uses of less expensive, lower technology contexts. We summarize in tables these 3 dimensions, which have been described earlier in the paper.

We also provide examples of effective combinations of the type of evaluation context, the type of materials used, and the type of task. These are intended to illustrate the design concepts.

Section 12: Conclusions

Findings

- We provide the Sensemaking Model of monitoring that characterizes monitoring as a broad activity of understanding the current, dynamic situation. This model identifies component skills and knowledge needed for monitoring and thus provides the target categories for training. In turn, evaluation of a training program should assess whether pilots acquired these skills and knowledge.
- A variety of measures, which we describe, should be used as they provide complementary information. Many of these measures are useful across a variety of test contexts and can be applied at multiple Levels of Evaluation.

- Both formative and summative evaluation are important; evaluation of component monitoring skills may be particularly important for formative evaluation to guide training revision; this approach is familiar in AQP (Advanced Qualification Program).
- We provide a general framework and toolkit that can be widely applied. It is also important to develop specific content and scenarios that appropriately sample the situations in which monitoring is carried out.

Recommendations

- Training conceptual understanding is critical for monitoring and assessing reasoning from this understanding is an important part of program assessment.
- Use multiple methods and situations as they provide complementary strengths. Assess both the components of monitoring and the integrated application of the components.
- Simulators provide particularly valuable assessment environment as they provide both control and realism.
- Consider using diagnostic assessment designs to test program effectiveness.
- Consider needs for staffing and staff training.

Further Considerations

- We provide a general framework for assessment design. This is a toolkit and a design approach to be instantiated within an airline, rather than a specific collection of cases, problems, and test items forming a particular evaluation plan. Detailed content and emphasis remain to be specified. While the evaluation principles will be of general use, details and emphasis in specific content will differ depending on operational setting, pattern of existing pilot skills, and specific goals of the training program to be evaluated.
- Capitalizing on existing resources for data collection within the organization (typically, an airline) is important and may be critical for reducing cost of a large-scale evaluation.

Future Work

- Development of training programs for monitoring and methods for evaluating such programs should develop hand in hand. The Sensemaking Model's analysis of monitoring skill can be used to analyze and build up this complex cognitive skill. Careful consideration will need to be given to how monitoring relates to other skills, and implications for training.
- Many useful measures of monitoring components and integrated monitoring performance exist at the level of impact on pilot performance and these can be tailored to the particular tasks and content needed. However, impact measures at the operational level deserve considerable development.
- Focused discussion with airlines should attempt to identify what situations or topics seem to pose the biggest monitoring challenges, identify what core content is common across airline operations, and explore the degree of commonality and difference across airlines and fleets.
- Development of training and of assessing that training "in miniature" for a bounded set of situations that are known to be important is a good next step. For example, applying the Sensemaking Model of monitoring, in detail, to monitoring from top-of-descent briefing with a particular focus on monitoring and managing complex Air

Traffic Control (ATC) clearances might be a focus for development. A matched program assessment for this limited domain could be developed in parallel.

- Developing a library of scenarios that pose particular monitoring challenges is an important resource for research and for operational use. These can be specified abstractly to allow implementation on a variety of fleets and also have accompanying cockpit video of pilots flying the scenarios and exhibiting some variety of monitoring behaviors.

1. Purpose and Scope of Report

Improved training of monitoring has been identified as an important route to improved safety and reduction of risk of loss of control (CAST, 2014; Flight Safety Foundation, 2014). Training programs aiming to improve skills and knowledge for monitoring will need to be evaluated. Evaluation of the skill of an individual pilot is part of program evaluation and, thus, many of the measures and methods we propose are informative at the individual pilot level. However, a broader set of considerations is needed for assessment at the program level. Our report provides a framework for assessment of programs that train pilot monitoring. Additional discussion of monitoring and a literature review focusing on training monitoring with secondary mention of training evaluation is provided in Mumaw, Billman, & Feary (2020). As used in training generally and in this report, the terms *evaluation* and *assessment* do not have a sharp contrast in meaning. *Evaluation* may emphasize the overall project, while *assessment* may emphasize the particular measurement methods used.

The need for improved training of pilot monitoring was motivated by inadequate pilot monitoring of aircraft state and flight path, which contributed to loss of control incidents and accidents (Mumaw, Billman, & Feary, 2019a). Our framework for program evaluation is inclusive enough, however, to cover training programs targeting pilot monitoring in general.

Our framework is based on a model of monitoring, which identifies the component skills and knowledge needed for effective monitoring performance. These skills and knowledge types specify the possible targets of training monitoring, and we cover how effectiveness of training these can be assessed.

Our report describes methods for measuring how effectively a training program met the various training goals. It describes a wide range of measures to assess training outcomes, suitable for the various components of monitoring, and applicable in various evaluation settings, from classroom to revenue flights. Various measures enable drawing operational tasks into controlled, diagnostic settings and also identifying measures suitable for operational settings.

The research reported here provides a framework for evaluating the effectiveness of training programs that aim to improve pilot monitoring. Our framework characterizes monitoring as a sensemaking process that builds a relevant, accurate model of the current situation. This framework provides an overall organizational system for selection and development of methods for evaluating training effectiveness. We do not provide lists of particular test items or scoring checklists. Rather, this framework provides guidance and organization within which coherent, efficient evaluations can be developed for specific application contexts. As attention to monitoring and monitoring failure has increased, so has interest in training of monitoring and in the effectiveness of training programs. Accurate assessment may be particularly valuable for new types of training programs. An evaluation should determine whether a training program that was intended to improve monitoring actually did so.

As attention to monitoring has increased, so has recognition that monitoring is far from a simple process, and far more than "where you point your eyes." Indeed, the broad set of skills and knowledge needed for monitoring seems to underlie much of what people informally refer to as "flying ahead of the plane" or even "good airmanship." Thus, training programs and their effective evaluation may necessarily be complex and will certainly be important.

The overall purpose of this report is to identify and describe methods for determining: a) whether a program for training monitoring in fact produces improvement and b) what component skills and knowledge the training does or does not improve.

The information provided in this report helps:

- characterize monitoring skills and knowledge in a form suitable for measurement
- identify relevant performance measures with complementary strengths and applicable in various settings
- describe how different types of questions about effectiveness of programs for training monitoring can be answered
- point out how evaluation of a training program can be designed to provide the best information about program effectiveness

2. Monitoring Competencies: Targets of Training

2.1. Our Approach to Identifying the Targets of Training: Monitoring Competencies

Key Point. Monitoring is a process of making sense of the current situation. We propose a model of the skills and knowledge needed for monitoring, the sensemaking model. This identifies core competencies and provides a framework for training evaluation.

We provide a qualitative model of monitoring, the Sensemaking Model, which identifies monitoring competencies, and in turn, guides development and identification of relevant performance measures. An earlier characterization of the model was presented in the Mumaw et al., 2020. That description cataloged specific, diverse knowledge and skills implicated in monitoring flight path, broadly defined. This report gives a higher-level account of how knowledge and skills combine to produce effective monitoring. The model identifies core competencies and prioritizes training these competencies, a perspective shared with Evidence Based Training (ICAO, 2013). This perspective provides a relatively high-level, coherent account for guiding both training and the evaluation of training programs.

Very broadly, expert monitoring depends on an accurate understanding of the current situation and its implications for action. An important training goal is to produce basic monitoring skills along with refinements to these that might otherwise only be gained with extensive experience. In turn, this builds piloting skill for adapting to the unexpected, as well as managing the routine, and accelerates adaptive expertise (Hoffman et al., 2013).

Identifying component competencies of the broader skill is important to provide specific targets for training. Our analysis of competencies is based on identifying the underlying cognitive processes and knowledge representations needed, summarized in our model. This analysis and model come from: a) a task analysis based on detailed conversations with and observations of expert pilots narrating what they are thinking while monitoring challenging situations, in combination with b) cognitive principles.

Comparing our approach with current AQP requirements shows that both divide complex work into components, which then serve as the basis for evaluation. AQP uses a Job Task Analysis (JTA) to identify tasks and subtasks (AC 120-54A FAA, 2017). JTA identifies Task Proficiency Objectives (TPOs) and Subtask Proficiency Objectives (SPOs). These tasks and subtasks are the goals of training and must be assessed in pilot evaluation. To do this, these tasks are mapped onto specific tests, behavior markers, and criteria to be used in assessing pilot performance. In addition, Enabling Objectives (EOs) are identified. EOs capture skills and knowledge needed for the tasks (and are assessed at some point in training), but, unlike TPOs and SPOs, proficiency may not need to be directly included in final qualification testing.

The analysis we provide is a higher-level framework than a particular JTA. At this level, we aim to identify core competencies needed for monitoring as sensemaking. This approach emphasizes competencies rather than tasks, as the same monitoring competency may show up in a considerable variety of tasks. Certainly, any performance measurement depends on identifying behaviors to observe and score. However, monitoring competences may include a mix of quite specific, task-like components and skills of very wide applicability that do not map clearly onto a specific set of tasks. For training, of course, exposure to instruction about broad competencies and practice with particular cases and tasks are both important.

Particular cases and tasks need to be specified to evaluate a specific training program. The airline would need to design the specific tests, behavioral markers or criteria most relevant to their operational needs. We do not define a particular test item to measure performance on a specific monitoring skill; rather, we provide a variety of measures (in Section 4) that could be used in assessment. The specific questions, items, or scenarios remain to be fleshed out in accord with program needs. Monitoring is an internal, cognitive activity and it pervades every high-level task in piloting; thus, there may not be a tight, logical mapping to a specific behavioral indicator but, rather, a variety of equally useful test items and cases. We have not traced out the details of how our framework aligns with the current AQP framework; integration with other pervasive, non-technical skills such as communication might be useful in developing an evaluation.

2.2. The Core of the Sensemaking Model: Pilot Situation Models

Key Point. The situation model represents the pilot's understanding of the current situation and is updated by a cycle of monitoring processes. The monitoring cycle forms a spiral where the situation model continually evolves.

The Sensemaking Model of monitoring identifies the major processes of monitoring and the resulting integrated understanding of the situation (Figure 1). We refer to this integrated understanding as a situation model. The pilot's situation model is at the heart of the monitoring model. The situation model integrates: 1) generally applicable mental models of "how things work" from long-term memory and 2) current information about the dynamic situation. Mental models provide a scaffold, frame, or schema for building the situation model and for interpreting events. While the boundaries among mental models are certainly not sharp, there is utility in proposing mental models of particular, tightly interrelated topics, such as the autoflight system, the National Airspace System (NAS), or what happens flying a Standard Terminal Arrival Route (STAR). Mental models can vary in completeness and accuracy, and they are modified by experience.

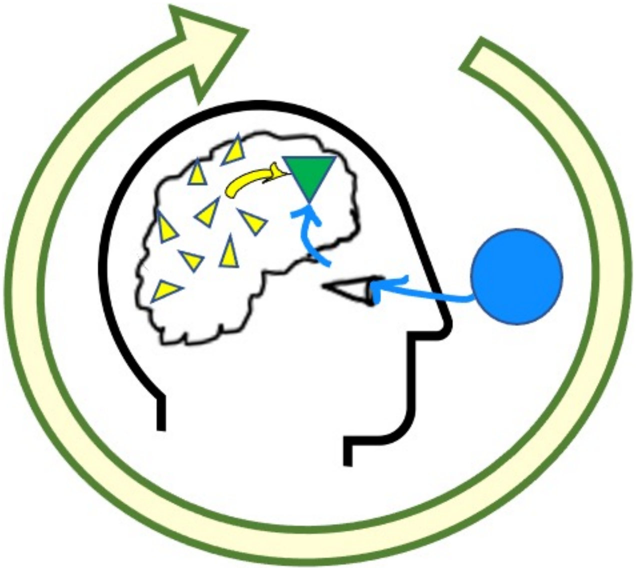

*Figure 1. The Sensemaking Model of monitoring claims effective monitoring is a cycle of selecting **mental models** (▶) and knowledge from memory and comparing that with information from the **world** (●) to form and update a **Situation Model** (▼).*

The dynamic situation model is initialized by selecting and tailoring a relevant mental model. This provides expectations about how this type of situation is likely to unfold in the current situation. The situation model guides monitoring processes and is, in turn, updated by those processes. The situation model not only includes understanding of the current state of the external situation, but also "meta information" about the situation model, such as what information is "old" and is due for an update, or a plan for ongoing monitoring.

The situation model is initialized with existing knowledge and updated through the cycle of monitoring activities: prioritizing what question needs to be answered, answering the question by gathering current information and comparing this to expectations, and identifying the implications for action. Successful execution of the monitoring cycle requires effective management of cognitive and physical tasks. Communication is needed to share content of the situation model and can also be critical in carrying out the monitoring cycle.

2.3. Domain Knowledge: Pilot Mental Models and Experience Support Monitoring

Key Point. A situation model is initialized with a) the pilot's background knowledge and b) initial observations about the current situation. Important background knowledge includes both organized mental models of how things work (also called frames or schema) and the pilot's general experiences and heuristics.

Useful monitoring depends not only on sensing current variable values, but on understanding their significance. This requires understanding what is normal, expected, and within limits for the observed variables in the current context. The foundation for these reference points are the mental models, which represent many types of systematic knowledge and procedures relevant to monitoring [Note that the terms mental model, frame, and schema have similar meanings]. The importance of mental models for effective thinking and acting is widely recognized (Gentner & Stevens, 1983; Kieras & Bovair, 1984). Mental models accessed from long-term memory provide an understanding of "how things work." Without such underlying knowledge, an effective model of the current situation—if the situation is complex—cannot be developed. If this knowledge is integrated with information about the current situation, it forms the situation model. In turn, this can be used to "simulate" or project what will happen in the future and explain how the current situation has developed in the past.

Additional knowledge from experience is helpful as well. Such knowledge includes familiar episodes or cases about how a problem was managed, both directly experienced and experienced through the narratives of others. This might include knowledge about how ATC manages traffic at a certain airport in particular conditions or how to configure displays to facilitate monitoring while using automation on approach. This knowledge includes a variety of heuristics such as the three-miles-to-descend-1000-feet rule of thumb (the 3-to-1 descent rule). Knowledge from experience may also update or elaborate mental models. Operational knowledge and mental models are needed to do the reasoning behind effective monitoring. These set expectations and guide comparisons, for example, comparing current with expected rate of descent. Further, pilot mental models and background knowledge can be incomplete in critical respects. Therefore, important objectives for training monitoring and for evaluation of the training program can be: a) improved background knowledge, particularly organized in mental models and b) improved skills to recall and apply the relevant mental models. For example, monitoring flight path may be improved by learning a more complete mental model of how different autoflight modes act in varied conditions; correct anticipation of likely ATC clearances on a particular STAR may be aided by experience flying the STAR with heavy traffic.

2.4. Initializing a Situation Model

Key Point. A situation model is initialized with a) the pilot's background knowledge and b) initial observations about the current situation. Important background knowledge includes both mental models of how things work and general experiences and heuristics.

Figure 2 illustrates a pilot's initialization of a situation model. A situation model is initialized with structured knowledge from a pilot's long-term memory, particularly their relevant mental models. Observations about the current context can be used to elaborate, set parameters, or modify the generic mental models to fit the current situation. For example, a pilot's mental model might specify the set of typical cruise altitudes for a cross country west-to-east flight, and a pilot's situation model would include the currently entered value of cruise altitude from the flight plan and allow comparison to assess whether this value is sensible.

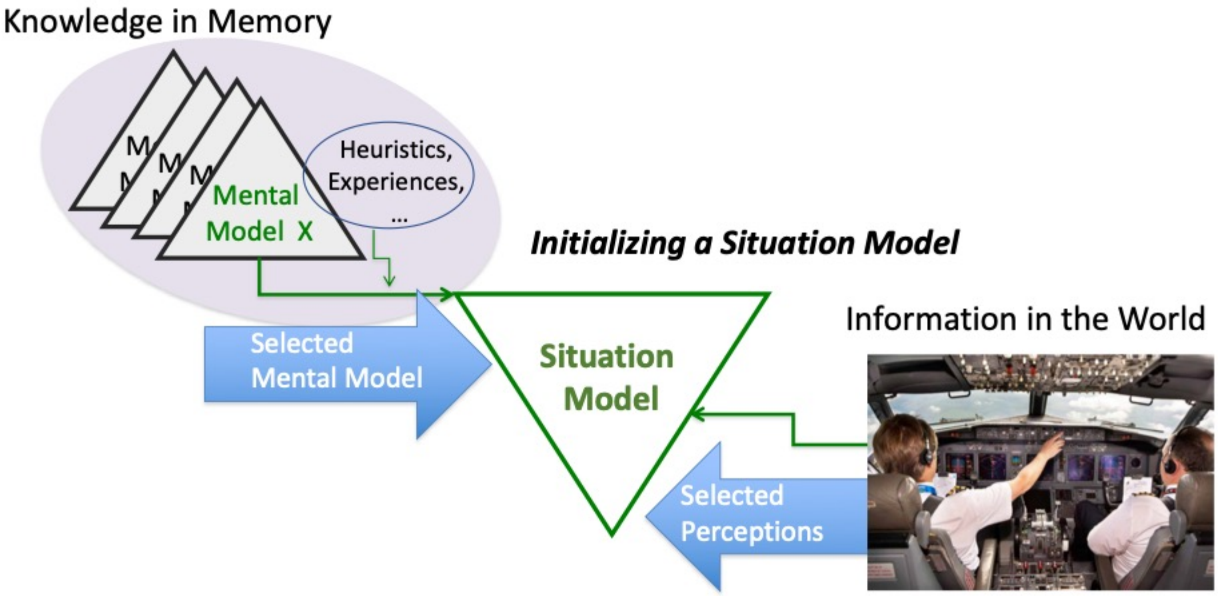

Figure 2. Initializing a situation model draws on knowledge retrieved from memory and information perceived in the world. Developing expectations from the situation model can guide selection of perceptions and perceptions can shape what mental models are retrieved as likely to be relevant.

2.5. The Monitoring Cycle

Key Point. A situation model is updated by a cycle (or spiral) of monitoring activities: identifying the specific, current aspect of the situation that needs to be checked or understood; finding and assessing the relevant evidence; and assessing the implications for pilot actions.

2.5.1. Focused Questioning

Figure 3 represents the formulation of a specific question or monitoring goal about the current or future state of the system, such as a need to update a critical parameter or check autoflight mode. It is not possible to check “everything” at once. Monitoring as sensemaking considers that effective monitoring is much more thoughtful than rote execution of a scan pattern. That is, the pilot must identify the information that is most important to gather or update in the current context. This might be a gap in the model, something puzzling, or stale information. Sometimes, framing a question will be cued by an external event, but for this to be effective, the meaning of the cue-in-context needs to be understood. Sometimes, picking the question will be familiar and routine, such as checking status prior to top-of-descent; in other contexts, this will require identifying critical gaps in the situation model and the variables needed to assess an emerging threat. Further, since attention is limited, key information in context needs to be identified and sampled appropriately.

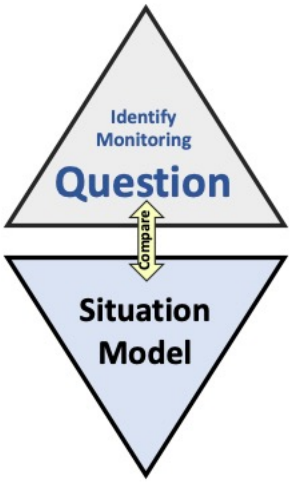

Figure 3. Updating the situation model. What is the most important question right now?

Forming a specific question, such as the monitoring goal, is important in order to make effective use of selective attention. The situation model provides the basis for identifying what question is most relevant to ask in the current situation and this is key for skilled monitoring. Monitoring questions might ask what the current value of a variable is, perhaps because a change is expected or because the value has not been attended to recently. Selecting what to attend to, based on the current situation is likely to be much more useful that a rote scan pattern or just relying on what happens to be noticed. Importantly, most useful questions ask not just what the value is of some variable but ask how an observed value compares to the normal, the planned, or an otherwise expected value. Questions might also ask about what a value should be and may involve consulting reference materials as well. New information from question-asking is added to the situation model and can also, of course, be added from casual noticing of values. Further, alerts are designed to pull attention to threats and the situation model can be modified this way as well.

2.5.2. Gathering the Evidence

The triangle on the right in Figure 4 represents gathering and assessing the evidence needed to answer the monitoring question. Different questions will require different information and different comparisons of that information. We use the term evidence because it is not just the data but the analysis of the data that produces evidence bearing on the question at hand; for example, comparing multiple separate sources of the same variable when the value on one source seems suspect. The situation model supports reasoning about the evidence and contains "meta-cognitive" processes and information about making sense of evidence. Collecting evidence requires knowing where needed variable values can be found, how to access them, and how to navigate to and configure displays. It requires understanding relevant thresholds, norms, currently required values, and how to make comparisons and projections using these data. A simple question—Am I at the cleared air speed?—might be answered by looking at the air speed window on the Mode Control Panel (MCP) for the cleared air speed, at the Primary Flight Display (PFD) for the current air speed, comparing these, recognizing the small fluctuations in current air speed are well within normal variation, and that the two values sufficiently match. Consider a more complex question: Will I make each waypoint constraint during descent given that there is an unexpected tailwind? An experiential rule-of-thumb may be helpful, but it may not be possible to project flight path precisely. Skilled pilots have a variety of heuristics for identifying what issues need on-going questioning and what data gathering and comparisons are appropriate for determining current status and the degree to which control inputs are needed.

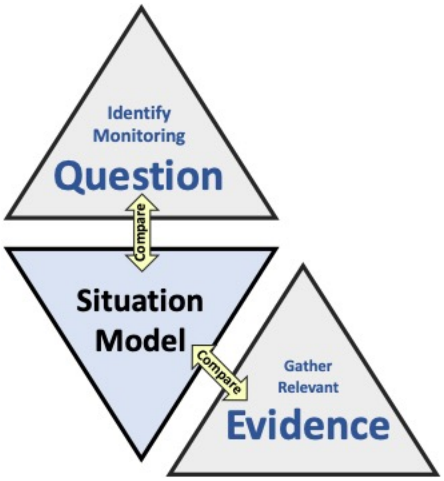

Figure 4. Updating the situation model. Given the current question, identify the relevant evidence to gather, compare, and evaluate.

2.5.3. Implications for Action

The triangle on the left in Figure 5 represents assessing implications for action. As the situation model is updated, it spurs consideration of the need for action. Needed actions may concern monitoring or control. Identified control actions and action choice may concern continuing under autopilot control with the current flight plan, reprogramming the flight plan, changing control mode or other manual control actions, or reporting unable to comply with a clearance. Actions to manage systems such as turning on anti-icing or changing airplane configuration may also be considered.

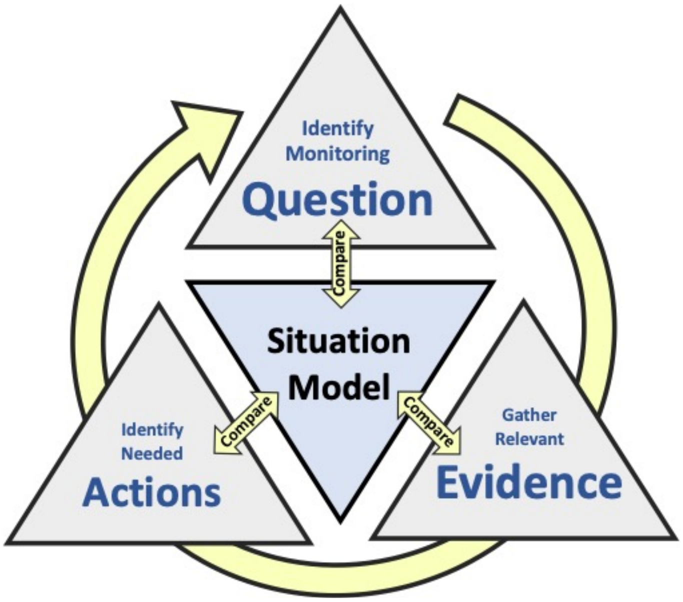

Figure 5. Updating the situation model: identify needed actions and within the interactive cycle of monitoring. The update spiral: a) identifies the question; b) gathers evidence; and c) identifies actions. Reference to the situation model guides the update processes..

Figure 5 also indicates the cyclic and interactive nature of processing. More exactly, monitoring creates spirals of activity updating the situation model. Each of the phases interacts with the knowledge in the situation models through processes of comparison and selection.

- The monitoring question is created or selected based on understanding what information is missing or stale in the situation model; in turn the situation model tracks what question is being posed and what sort of answer is expected.
- Effective evidence-gathering requires understanding what information will answer the question, where to find the information and how to analyze the information. Analysis may include comparison of current values to values predicted by the situation model, comparison of current values to normal values or to values required for the type of situation, or consistency checks within the information gathered.
- Action evaluation is fundamentally concerned with assessing whether and what future actions are needed. For flight path management, this means assessing whether the current control modes will produce the desired flight path over the next interval, but actions can also concern tests or checks to carry out.

For simplicity, we label all the ways these processes interact with and use the situation model as “comparison.”

Here we briefly place this model in context. This model is a simplification of all the processes involved in monitoring. For example, “bottom up” processes of noticing unexpected information also contribute to monitoring but are not highlighted in the model. The model presented here focuses on the activities we think are most relevant to training. The Sensemaking Model of monitoring emphasizes the cyclical and interactive relations among processes and the pilot’s situation model. Some models of complex cognition and complex work emphasize a linear beginning-to-end structure. A comparison between this cyclic model and a similar linear model is given in Appendix A.

2.6. Attention Management and Communication

Key Point. Monitoring takes place in a complex task and social context. The pilot needs to manage multiple tasks effectively, keeping monitoring suitably prioritized. The situation model needs to be communicated, and communication also provides input and guidance to the process of monitoring.

The monitoring cycle takes place in a complex context, as illustrated in Figure 6. The green area on the left represents events in the world that make demands on attention. The blue area on the right represents a pilot’s mental work, or cognitive processing of other tasks. The grey rectangles illustrate other mission-relevant tasks, while the tan rectangles represent irrelevant, distracting activities. The pilot must regulate attention so that monitoring takes appropriate priority. Attention needs to be regulated to resist distraction from internal or external sources while remaining sensitive to important interruptions. Tasks need to be managed so that other necessary tasks are accomplished while monitoring is maintained. Some of these regulatory processes in task management occur at an individual level, and some concern crew-level coordination.

Crew communication is critical to monitoring as well as for task management. Communicating about the situation model is important for both Pilot Flying (PF) and Pilot Monitoring (PM) roles, but the communication may be a higher portion of responsibilities for the PM. Clearly, the PF as well as the PM needs to have a relevant, accurate model. Situation models do not need to be completely shared—the two pilots are unlikely to be monitoring the same indications—but their models do need to be consistent. Communicating between the pilots is critical for maintaining consistency.

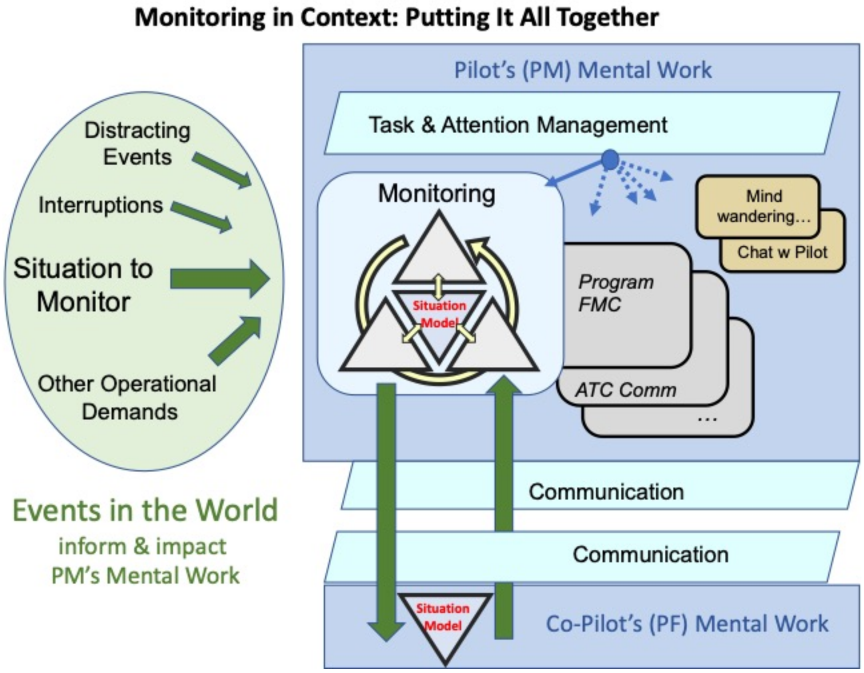

Figure 6. Monitoring depends on Task and Attention Management and on Communication. Monitoring takes places in the context of other tasks and events thus requiring regulation of attention. Communication is needed both for building and sharing the Situation Model.

As described in more detail in Appendix A, cognitive models can emphasize a linear sequence of steps from perception to action. This type of model may be more familiar, and it can be applied to monitoring, as shown in Figure 7. This shows a sequence of steps and also indicates which components are part of monitoring. The Sensemaking Model of monitoring includes but reorganizes these components, emphasizing the pervasive role of comparison and situation understanding throughout the processes of monitoring. Figure 7 provides a rough illustration of how linear processes may be included in the cyclic model.

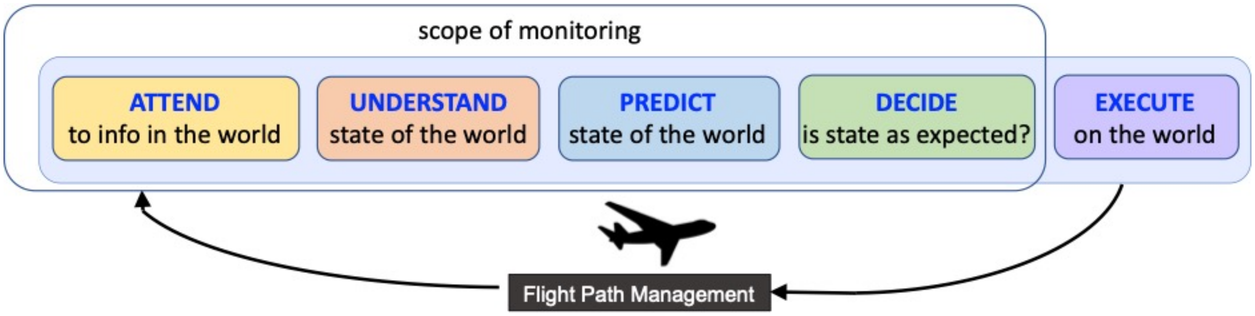

Figure 7. Linear models emphasize a sequence of steps from attending through thinking and action. Here, we consider that monitoring includes all the illustrated processes except execution of action. The linear structure predominates even if feedback is acknowledged.

2.7. Monitoring Boundaries and Integration

Key Point. Monitoring skills are related to and overlap with other CRM or Nontechnical Skills. Training and evaluation of monitoring skills can draw on these relationships.

Monitoring is an encompassing activity, and components of monitoring are also components of other training units. Indeed, monitoring is one of a set of ‘cross-listed’ topics that include elements of CRM (Crew Resource Management) or the NOTECHS (NonTECHnical Skills) skill clusters (e.g., situation awareness, communication; see Flin et al., 2002), or other related programs: TEM (threat and error management) and VVM (verbalize, verify, monitor); and autoflight system knowledge. A key outcome of successful monitoring training will be effective integration of monitoring into these other skills and knowledge to support overall operations. How introduction of an additional training unit for monitoring affects and is affected by existing programs on related topics is an important factor to consider in training and evaluation. A relevant area of training assessment is how well the monitoring training is related to other concepts; for example, are concepts confused or are the relations between concepts understood. Clear understanding of terms can be helpful, both because this facilitates communication and because terms and concepts act as retrieval cues for recognizing a situation or identifying relevant actions to consider.

This monitoring model applies to all monitoring topics: monitoring of flight path, of systems, of crew, etc. However, the initial motivation and evidence for the model came from flight path management.

3. Training Program Evaluation Levels

Training evaluations can answer several types of question and serve a variety of goals. Evaluation may focus on long-term, large-scale improvement at the operational level or immediate, fine-grained impact on the individual trained.

Kirkpatrick labeled four evaluation levels distinguishing what outcome question was evaluated, the associated kind of measures used, and where the measurements were taken (Kirkpatrick & Kirkpatrick, 2006; Kirkpatrick, 1956). This was originally concerned with training for management domains quite different from aviation and monitoring. Level 1 evaluation measured trainee attitudes about the training, typically with a survey provided in the classroom environment. Level 2 evaluation measured whether trainees learned the concepts and skills that had been taught, typically measured with open-response or multiple-choice tests. Level 3 evaluation measured individual trainee’s performance at work. Level 4 evaluation measured organizational metrics such as profit, new accounts, or reduced waste. For many of Kirkpatrick’s domains, it was only possible to do anything close to actual work (e.g., a salesman trying to land a new customer) in the actual work context.

Modern training and testing tools for aviation (and other domains) provide slightly different opportunities. More interactive and realistic activities can be presented in the classroom and simulator for both training and program evaluation. Highly realistic work can be assessed in training facilities as well as in actual flight. This led us to split the work performance Level 3 into two, based on whether work performance is measured in the simulator versus in revenue flight (both originally Level 3). For aviation, levels of evaluation can be characterized in this way:

- Level 1. Measures pilot’s opinions or attitudes about the training, often assessed in the classroom or other non-operational environment. Level 1 evaluation is not generally concerned with assessing performance or whether the training produced change.

- Level 2. Measures mastery of the component skills and knowledge underlying pilot performance that was the target of training, often assessed in the non-operational environment of the classroom.
- Level 3A. Measures pilot or crew performance in a simulator that has a fidelity level appropriate to the competencies that are being evaluated.
- Level 3B. Measures pilot or crew performance during actual flights.
- Level 4. Measures an organizational performance metric, generally a metric of operational safety or efficiency.

Levels 2–4 make use of the pilot performance measures described in Section 3 and add others. These levels are concerned with measuring change and estimating whether the change is produced by training. We discuss possible designs for testing the impact of training in Section 10. Our discussion of training evaluation methods is organized from lower to higher levels.

Measurement is generally easiest and least expensive at lower levels; it may also be easiest to identify impact on specific skills and knowledge. Targeted, lower-level measurement of specific components may be valuable for diagnosing where training was not effective and for identifying where higher levels of training should be focused. It is often important to know whether the training did not change the targeted skills and knowledge, or whether targeted skills and knowledge were not a critical limiting factor for performance. Measurement at higher levels may be most important for policy decisions, such as whether the training program should be retained and adopted more widely.

Training evaluation benefits from coordinated design of measurement at multiple levels. It is also useful if the program evaluation plan can be coordinated with evaluations conducted as part of the training and with data collection across the airline. For example, surveys asking participants about what aspects of training were strong or weak are better done while memory for the training events is recent. Further, such surveys may be valuable for guiding training modification. Thus, this aspect of evaluation might be conducted as part of the delivery of training.

4. Measures of Monitoring Performance

4.1. The Importance of Measuring Performance

A training program is usually motivated by the goal of improving fleet operations in some respect, such as improving fleet safety. However, the most common measures of fleet safety—accidents and incidents—are rare and, therefore, are not sensitive measures. Thus, a range of performance measures and tasks are needed. Performance measures relevant to monitoring, in the context of interface evaluation are also discussed in Billman, Mumaw, and Feary, 2019.

Assessment of monitoring is complicated because much of monitoring is cognitive, and not a matter of highly visible actions as in manual control. At a very broad level, the primary types of performance measures for assessing monitoring effectiveness concern what the pilot says and what information seeking or control actions the pilot takes. Monitoring very much concerns awareness and intent; asking for a verbal report is typically a very efficient way of getting information. Talking about what is in awareness or what the pilot is trying to do usually is a highly practiced activity that does not interfere with the aviation tasks; evaluators should, however, be alert to any evidence of exception to this pattern. In addition, verbal-reports about what the pilot knows about some topic, such as autoflight modes, can be valuable. Nonverbal behaviors are informative if they are intentional, discriminative actions; that is, they would only and reliably be taken in a specific context, if and only if the pilot had a particular awareness or intent. These verbal and nonverbal intentional, discriminative actions (IDAs) can be assessed for quality and for speed.

Information about where the pilot is looking, and where not looking, may provide supplemental information about monitoring. A variety of physiological measures are under investigation. However, we suspect these measures are not linked directly enough to monitoring, awareness, or understanding to use them in evaluation of training programs. A distinct though related question about the use of eye-tracking concerns whether providing such data for an individual pilot to their trainer/evaluator improves the trainer's ability to assess the individual's monitoring skill.

4.2. The Strategies and Issues for Measuring Performance

Key Point. Several issues or strategies influence the usefulness of performance measurement. These include realism vs diagnosticity of evaluation context, variation in difficulty of items or tasks, assessment of a factor when that factor is limiting performance, measuring generalization, and diagnosticity of scenarios. Use of multiple measures with complementary strengths is valuable.

4.2.1. Use Multiple Types of Evaluation Context

Many of the same aspects of performance can be measured in different evaluation environments, but simple versus complex environments have different benefits. Simpler contexts in more constrained environments allow more diagnostic and faster assessment of specific aspects of competence. In contrast, if performance is poor in a complex context such as a revenue flight, it may be very hard to diagnose what competencies are lacking. Further, while the complex environments of actual flight allow exposure to the full work demands, they may not provide the opportunities to assess specific skills. For example, opportunities to practice management of emergencies in actual flight are extraordinarily rare. Active learning environments, and particularly simulators, also provide powerful assessment tools where task context and duration can be varied from animated "snippets" to flight-length practice. Usually, measurement in multiple contexts and levels is valuable in assessing effectiveness of a training program. Importantly, previously collected data may also be available as part of the training program itself and can be used for program evaluation.

4.2.2. Use Items that Vary in Difficulty

Because individual pilots vary in their pattern and level of competencies, having assessment tasks at different levels of difficulty is informative. There can be exceptions to this rule of thumb; if there is a sharp "pass" criterion and all that matters is assessing whether performance falls above or below this point, then testing concentrated around the criterion may be appropriate. However, we doubt that criteria of this sort well defined for monitoring skill. Assessment may be most effective if tasks vary in difficulty. For example, assessment tasks and scenarios to monitor can vary in the type of challenge (or variance), such as weather, terrain, crew and in how hard a problem is to recognize and understand.

4.2.3. Use Conditions Where Performance Actually Depends on the Target Competence

To test a specific competence, performance on the task needs to be higher (or lower) when that competence is higher (or lower). If some different factor is what is limiting performance, the task won't be measuring the target competence. For example, if a warning in a certain task cannot be read, the task won't be measuring how well the message is understood and if the goal is assessing overall monitoring skill in a simulator, performance should not be limited by the requirement for highly expert manual flying skills. Similarly, if a measure of awareness requires that a pilot report something at the end of the event, performance is likely limited by forgetting and not by what was noticed in the moment. In addition to being specific to the target competence, performance on the measure should not be too high or too low; "ceiling" and "floor" effects should be avoided. If virtually all pilots are already scoring very well on a particular measure before training, there is no room to measure improvement; the task as measured this way is too easy. Perhaps the measure can be changed; for example, using speed rather than correctness, if speed is actually important for the task. Conversely, if the task is very difficult it may not reflect improvement, even if underlying skills did develop. Unrelated contextual factors can be used to modulate task difficulty while still ensuring sensitivity to the target skill; for example, tasks assessing whether critical changes are noticed can be made more difficult by adding potentially distracting tasks that must be managed, thus producing high workload.

4.2.4. Consider the Similarity between Activities in Training and in Program Assessment

Evaluation of training programs should assess both how well pilots learned specific monitoring strategies taught for specific types of situations and how well pilots monitor in situations for which they were not specifically trained. Proficiency is important both in situations with established, good monitoring strategies, and in unfamiliar situations. Obviously, not all situations can be included in training, and the ability to generalize to a wider range of situations is important to assess (Koteskey et al., 2019).

4.2.5. Scenario Design is Critical Across Measures

The task and situation in which behaviors are observed are central to the effective use of these measures. We refer to the task-situation combination as a scenario (Mumaw et al., 2019a) and use this term very broadly, applying it to classroom exercises and to events in revenue flight. "Event sets" in AQP-based design of line operational simulation (LOS) events correspond to diagnostic parts of scenarios in the context of simulator-based pilot evaluation. From the perspective of training program evaluation, the goal is to assess competencies across a strategically selected, broad sampling of scenarios. These will differ depending on the aspect of monitoring performance to be assessed.

4.2.6. Open Response and Closed Response have Complementary Strengths

There are tradeoffs in how easy it is to collect and score a task and how informative it may be. Generative (or open response) tasks are often more informative but take more time to do and are harder to score than choice (or closed response) tasks. Generative tasks are tasks like recall of studied material or predicting a sequence of events. Choice tasks are tasks like recognition of whether or not a studied fact is true or which of two explanations is correct. Control tasks may act like choice tasks if the action space is small and well-defined, such as choice between two autoflight modes; alternatively, if the action space is not well-defined or is large, the pilot must generate the needed behavior not just recognize which of two (or a few) alternatives is correct. Generative tasks are particularly helpful when developing items or when researchers are unsure what aspects, situations, and activities will be easy or hard. To be useful, choice tasks have to well calibrated to what pilots do and do not know.

4.3. Program Effectiveness: Training of Mental Models and Heuristics

General operational knowledge and knowledge held in mental models are critical to develop a situation model. Effective monitoring depends on understanding how things work. This enables explaining what has happened, projecting to what will happen in the future, and identifying relevant actions. Verbal explanations provide one type of observable behavior that can be measured and compared between pilots, with and without the training being evaluated.

Several types of knowledge contribute to understanding. Mental models provide an organized, if incomplete, causal model of some aspect of the domain. Domain expertise depends on mental models for how systems within the airplane work, flight dynamics of the airplane as a whole, operation in the airspace, weather, and more. Assessing the content and use of mental models can be an important evaluation goal. In addition to mental models, expert monitoring draws on a variety of heuristics—such as the three-to-one rule for descent rate, advice from other pilots, and memory for particular arrivals. These may be acquired informally or in training. It is not feasible to assess all the important, relevant knowledge and skill. Therefore, evaluation should target those topics most important for the training program to improve, perhaps because pilots show greatest deficit on these topics or because they are new.

4.3.1. Measure of Fact Recall or Recognition

Memory for basic components of knowledge can be assessed through recognition (closed response, choice task) or recall (open response, generative task). For example, a multiple-choice test could provide a recognition test of factual knowledge about how the autoflight system works. Recognition tasks have the advantage of being very easy to score and can be designed to require a pilot to discriminate the correct answer from other plausible choices. A recall test could ask for pilots to write down basic facts about how a mountain range affects wind pattern, or what the rule of thumb is for estimating feasible descent. Recall tasks require more of the pilot and often are more realistic but are also harder to score.

4.3.2. Measure of Explaining a Situation

Generating an explanatory text or diagram requires coordinating and relating multiple facts, and thus tests a deeper understanding than recall or recognition. An explanation task may be framed as teaching something to an inexperienced pilot. Often, it is useful to request relatively constrained explanations or to explain a particular contrast or difference. A test might ask how difference in performance or airplane behavior is affected by differences between two vertical autoflight modes, or to explain the rationale for how ATC manages traffic for merging STARS with different traffic levels. Concept maps are one way to ask for the content of a mental model, but concept maps are usually unfamiliar and very unconstrained; asking for explanations seems to be a more natural method. In addition to providing explanations that draw on general mental models, pilots can also be given a situation and asked to "explain why this happened" or why some state occurred. The situation may be provided in a description, a series of displays, an animation, or a video. The situation may also be taken from a simulator or actual flight and the pilot asked to explain a specific aspect or situation in debriefing. This task capitalizes on the familiar activity of debriefing but engages that activity on a (set of) standard situations. The pilot might be asked to "explain" the whole event (as might happen in a normal training debrief) or might be asked to explain a specific trigger or planned aspect (for example, the effects of a mode change in a particular context).

4.3.3. Measure of Prediction from a Situation

Explanation and prediction are closely related but prediction tasks focus on 'what will happen next' from a certain type of situation. Both processes are important to monitoring. Switch in perspective from what has happened to what will happen is useful. If a pilot has explained a past event incorrectly, this will easily lead to inaccurate projection of what can happen next, as well as prompting the wrong questions. Focusing just on what happens immediately next will make the task more constrained. Predictions can be asked about the overall situation or about particular variables. Prediction tasks can be framed about general types of situations and draw on a mental model. An example of a general prediction question is here: If you are never cleared to the cruise altitude in the flight plan and without pilot intervention, what autoflight vertical modes will you sequence through from climb to approach.

Predictions can be very specific and draw on the current situation model, as here: If your flight plan specifies a cruise altitude of FL330 and top of descent occurs at XX waypoint, but you were never cleared above FL310, what mode will you be at XX without pilot control action?

4.3.4. Measuring Model Selection

Selecting relevant mental models to initialize (or update) a situation model, building a related situation model, and recognizing a mismatch between the current situation and the underlying mental model (or frame) are all important parts of expertise. Effective monitoring depends both on activating a mental model that is relevant to the current situation and recognizing if the underlying mental model no longer matches the current situation. Failure on any component in the monitoring cycle may be caused by building the situation model on an inappropriate mental model. Measuring the model selection process is difficult to assess but may consist of two phases: a) recognition that the current frame for creating expectations is amiss and b) adoption of a better matching alternative frame, or perhaps the "meta frame" of "Identify what is going on." Startle and surprise can indicate recognition that the current model (or "frame") does not fit, and the subsequent reinterpretation may reveal the process of model selection. Startle and surprise have been measured in simulator studies. There are (at least) two distinct paths for producing surprise. First, a situation that is in principle highly predictable might surprise a pilot with a deficient situation model; the model might not include the information needed to make the prediction (Sarter & Woods, 1995). For example, unexpected failure to descend might reveal problems in the pilot's mental model of autoflight systems or of how they are engaged in the current situation. Second, the external situation might unfold unpredictably in ways that change what model is relevant; for example, weather conditions or equipment failures might require rejection of the active model and replacing with a model that better fits the changed conditions. As research on startle and surprise progresses (Landman et al., 2018), measures of startle, surprise, and recovery will be important performance measures of monitoring.

4.4. Program Effectiveness: Training Monitoring Proficiency throughout the Monitoring Cycle

We have labeled the model updating process as the monitoring cycle, but more accurately it should be thought of as a spiral; that is, while the general types of processing repeat, the situation model evolves with each loop. Measuring the specific parts of the monitoring cycle that are strong or weak across different situations can be helpful, as this can diagnose specific weaknesses in the training program. Training effectiveness can be evaluated for each phase of the monitoring cycle. Successful application can be identified and gaps or limitations, as listed here:

1. Is the pilot addressing a low priority question?
2. Is the pilot unsure of what evidence about the current situation is needed to answer the question, of where the information can be found, or of what reference values and comparisons should be used to understand the unfolding events?
3. Has the pilot failed to identify the need for pilot intervention, misidentified the action needed, or concluded an action is needed when none is currently called for?

4.4.1. Identifying the Monitoring Question: What Monitoring Question When

How well did the training program teach pilots to pick the important monitoring question for the immediate context? The same specific issues do not need to be understood in all situations. For example, if ATC just issued a clearance modifying the STAR the pilot is currently flying, it is relevant to assess whether the pilot can safely and certainly make the new clearance and also the next restrictions on the cleared flight path. Some questions are always relevant but not with the same frequency or priority relative to other questions; for example, assessing whether the airplane trajectory respects clearance tolerances needs to be done more frequently in dynamic phases of flight than in cruise. For some situations, relevant question strategies can be identified and trained for formulating relevant monitoring questions. We consider how to design flight scenarios that are useful for evaluating training and how to obtain data about how well pilots were trained to formulate monitoring questions.

4.4.1.1. Flight Scenarios

The flight scenarios used in assessment need to be selected strategically. They should prioritize specific monitoring questions so an explicit, objective scoring method can be developed to assess questions asked. Particular training programs may train different question forms and content. For example, in each of a set of recurring situations, pilots might be trained to ask a particular question, and this may help the pilot retrieve the useful assessment activities. Most of the situations in assessment should be analogous to the situations used in training. However, it is also valuable to include some novel situations to assess pilots' ability to generalize.

4.4.1.2. Response: Verbalize the Question

Pilot understanding of how to focus on relevant monitoring questions can be measured by eliciting a verbal report across a variety of contrasting situations. The pilot, in an operational context, may be asked a direct question by an observer or instructor, for example, "What do you want to know or understand right now?" Operational contexts for a verbal report might be: 1) you are returning to the flight deck (in a specific context), what do you want to know first?; 2) you are coaching a junior pilot about what to attend to in this particular challenging situation; or 3) you watch a video of a flight crew and are asked to identify what they are trying to find out and whether this is the priority in this context.

Structured coding schemes are useful for open-response items, analogous to grade sheets for scoring CRM (Koteskey et al., 2019). Written or verbal responses can be used. Verbal responses are easier to produce than written but may be harder to score. The scenarios being assessed should have known correct components, and the primary scoring should code for whether these are present. An auxiliary score may be useful for coding inclusion of incorrect and/or low priority questions (e.g., in addition to correct questions) if pilots are likely to produce multiple or multi-part questions.

In addition to scoring for content, it can be useful to score for form, particularly a question's level of generality. Questions of intermediate generality, such as "Can we meet the ATC clearance," "Is heavy traffic likely to impact our ability to maintain separation," or "Is the PF using the planned autoflight modes" may be most useful. These questions invite the pilot to notice, compare, and assess multiple variables and produce a mission-relevant assessment. Intermediate levels encourage flexibility in how the question will be answered, thus supporting adaptability to conditions, but still provide structure and may cue learned monitoring patterns. Highly general questions are unlikely to be very useful: "I'm checking that we're ok" does not constrain what information should be gathered, compared, or provide any boundaries for when the question is sufficiently answered.

Highly specific questions, such as “What is my altitude” or “What was my last clearance” may lead to fragmented awareness not integrated into the situation model in a meaningful way. Put differently, specific questions such as these may be most useful as subgoals or parts of a broader, integrative question. In simple tasks, when people characterize what they are trying to do at a low level, they are more prone to distraction or disruption in accomplishing the goal (Vallacher & Wegner, 1987). The same principle may well hold for more complex activities.

4.4.1.3. Response: Select the Question

Asking pilots to verbalize their question is likely the most informative measure of skill, and having pilots generate rather than recognize a correct response is a better measure of training effectiveness. However, for particular trained situations with expected “right answers” it may also be possible to investigate pilot competence with multiple-choice tests. For example, a pilot might view a short video or “snippet” sampling a few minutes of flight, then be asked “What is the priority question to answer now?” and make a selection from several multiple-choice alternatives. Contingent on the response, the pilot might also be asked to identify why this is the priority from several alternative explanations.

4.4.2. Gathering Relevant Evidence

Gathering and interpreting evidence is at the heart of monitoring. The relevant evidence needs to be gathered and assessed to answer the active question and to address information that is missing, inconsistent, or unexpected in the situation model. Good performance requires knowing where to find data about the current situation to answer the question and how to assess the implication of data.

4.4.2.1. Verbal Report

As with measures of question asking, evidence processing can be assessed by verbal or written report. Verbal report is particularly helpful for assessing the comparison processes. Thus, all the same types of measures useful for assessing question-asking can be used for assessing evidence. Some additional verbal formats are available as well.

Assessment typically involves comparing current values to expectations, which may be based on an understanding of the current setting. Assessing pilot expectations for a particular situation can be probed verbally; for example, the pilot is introduced to a situation where partial information is provided and then asked what they expect the value of a particular variable(s) will be. Pilots can also be asked to note or to mark which variables are as they expected and which are surprising or unexpected. The pilot can be provided not only with a description of the situation but also of a good, target question to answer as part of setting up the pilot’s assessment task. The response here can separate ability to find and interpret information about the current situation from the ability to pose the more valuable questions.

4.4.2.2. Verbal Critique

In addition to demonstrating the process of evidence gathering, the pilot can be introduced to a situation and then given a description of how another pilot answered a monitoring question. The performance task here is to critique, improve, or provide feedback on what the “junior pilot” did. This reduces the task of producing the full account of evidence-gathering and relies more on recognizing good or poor approaches. Often it is easier to comment than to do.

4.4.2.3. Nonverbal Actions

Some measures of nonverbal actions are useful here as well, particularly for the data-inspection aspect. The pilot can be asked to select (point to or click on) the information sources judged most important to look at, given the current context; values may be hidden throughout or revealed when an indicator or other information source is selected by the pilot. This can be done in simple evaluation software that displays a sequence of flight deck states, and automatically records pilot-clicks on different indications in response to the situation or to specific questions. It can be done in a simulator or part-task trainer where pilot choices are recorded by an observer.

4.4.2.4. Eye-Tracking Data

Across the overall process of monitoring, eye-tracking data may be most useful in assessing data acquisition. However, the relation between eye fixation and awareness is complex (e.g., Mumaw et al., 2000). One possibly useful measure is the time since a particular variable was fixated (maybe for some minimum period of time); that is, how long since the value was sampled. Data recency sets a bound on when the pilot last observed the value, so long as the displayed value has to be fixated to be understood. Failing to sample important variables with appropriate frequency could be one measure of data gathering competencies. For scoring, here as elsewhere, specific scoring of response quality or correctness is important; for example, does a continuous measure of data recency matter or only whether a data source was fixated more recently than some threshold. Eye-tracking data may be helpful in training as an aid to instructors. While future research may provide informative measures using eye tracking, it may also be that gaze is too indirectly linked to understanding to be a good measure. Research on eye tracking does not yet provide much of direct use for assessment of training programs.

4.4.3. Identifying Implications for Action

The third phase of the monitoring cycle is identification of what actions should be taken in the current situation, particularly when this action depends on understanding the situation and not simply “rote” execution of a procedure. Measurement of correct and timely identification of context-relevant actions is an important aspect of assessing monitoring. Measurement can assess whether, and at what point, a needed action is taken or mentioned. Good measurement of monitoring usually depends on scenarios where recognition of some aspect of the situation requires a distinctive, intentional action, that is, the action would only be taken if this aspect was recognized.

In an evaluation context, the PM, PF, or crew as a whole can be the focus of assessment. If the PM is the focus, the time to first state the need for action, and time to state what the needed action is, may each be useful measures. The clarity and specificity of communication may be important to score as well. If the PF or crew as a whole is the focus, execution of the needed action is likely the primary measure, and actual execution of the action by the PF can be a useful variable for evaluation of the PM, as well. Efficiency of communication by the crew might be estimated by the time between first comment that implies a needed action, and the initiation of a control action (e.g., first change on the throttle).

Elements of the evaluation methods are much the same as for other components of the monitoring spiral. Scenarios should be selected that are relevant to the situations addressed in the training program. Specifically, objective scoring for both communications and control actions needs to be developed. Pilots can be asked to generate the solution or critique examples of actions identified in test examples.

In many situations, pilots are not under extreme time pressure, having minutes to assess and act. Some situations do require that the correct action is identified and executed very quickly, for example, go/no-go decisions on takeoff and, to a lesser degree, go-around decisions on approach. Emergencies from equipment failures (e.g., loss of pressurization) require quick responding but are heavily supported by alerts and thus may place less burden on monitoring skills.

A key element of competence is fast, fluent recognition that the situation requires a specific response. Training for rapidly unfolding situations should train fast responding, and evaluation of training program adequacy will similarly benefit from fast-paced assessments. For example, if the program trained pilots to quickly recognize no-go situations, speeded recognition could be used in program assessment.

While need for rapid response poses particular challenges to monitoring, slowly evolving situations pose distinctive challenges as well. Slowly developing changes may be hard to notice or understand. Slowly changing situations may particularly depend on appropriate management of the monitoring activity and of rotating through the range of evidence gathering to cover widely applicable, as well as situation specific, events unfolding at quite different rates.

4.5. Program Effectiveness: Training Task Management and Communication for Monitoring

The monitoring cycle, or spiral, that we described takes place within a broader scope of activity. How much of the related work is included in the training program may vary considerably, and program evaluation should adjust accordingly. Monitoring overlaps with several topics that are included in other training programs, including TEM and CRM.

4.5.1. Task Management

Task management is an important skill that impacts monitoring. An important element of task management is off-loading tasks that can be performed at a later time to ensure that non-essential non-monitoring tasks are not competing for the PM's attention. In addition to managing pilot workload, task management includes scheduling task flow so tasks, including monitoring tasks, are done at the most informative or useful time and are least vulnerable to interruption.

Task management, in relation to monitoring, can be measured in several ways. Advance planning can be assessed in how tasks are anticipated in briefings, such as in pre-flight or approach briefings, and whether activities are executed in accord with the plan (in the absence of unexpected disruptions). Task management can also be assessed by how well unexpected demands are managed when they threaten the ability of the PM to monitor effectively. These planning and execution elements can be assessed both for conceptual knowledge of what should be done and for the ability to execute under pressure.

Conceptual knowledge can be assessed by having a pilot formulate or critique a briefing that, without time pressure, includes task scheduling to protect monitoring; that is, a static flight context can be set up and available for reference so the pilot does not need to manage a dynamic situation. Conceptual knowledge can also be assessed at a crew level where role-playing (or hypothetical planning) can be carried out.

To assess execution under pressure, either the planning process or plan execution can be disrupted by additional, unexpected tasks as the pilots fly simulated scenarios. The scenarios can be designed so that the occurrence and nature of disruptions are hard to anticipate, and the disruptions create high workload. Scenarios should also be designed so key aspects of performance can be unambiguously scored, for example, it is clear what tasks should be shed (e.g., planning next leg, progress reporting), off-loaded (e.g., to PF or dispatch), or negotiated (e.g., amendment to an infeasible flight path clearance from ATC). One of the challenges in designing simulation scenarios is management of pilot expectations. Pilots expect that simulations will include failures and problems of various sorts, and thus there is likely greater anticipation of high workload and problem management than during revenue flights. One strategy is to create an expectation of a different type of problem, so that the nature of the specific challenge is not anticipated. In these activities, the assessment is targeting the training program, not the pilot. Thus, there may be greater flexibility in the range of acceptable situations. Such assessment might be carried out in the context of or in the time allocated for First Look assessment. (The First Look portion of AQP training is a no-jeopardy setting with which pilots are already familiar and that assesses proficiency at the fleet rather than individual level.)

As just discussed, task management, at one level, coordinates monitoring with other operational activities. Task management, or 'meta-monitoring,' also plays out at a lower level within monitoring. The monitoring focus is largely driven by the situation model, but priorities are also driven by external events and the passage of time. External events may draw the pilot's attention or "be noticed," update the situation model, and guide subsequent monitoring cycles. The passage of time means that it again becomes time to check basic parameters that may not have particular urgency but need to be checked with some regularity. Some of these regulatory processes may improve with experience or practice, such as increasing the salience of useful cues. Some situations with rapid classification based on predictable cues may benefit from training, as suggested for go/no-go decisions. However, we do not know of "generic" attention training that would guard against occasional slips or lapses. We suspect that the best protection is strategic, meaningful selection of relatively short monitoring cycles, interspersed with time to consider what is most important to review next.

4.5.2. Communication

Just as monitoring takes place within a broader set of flight activities, it also takes place within the communication context. Communication is a vital part of monitoring and monitoring effectiveness can be limited by inadequate communication as well as by inadequate understanding. Communication skill can appropriately be included in training for monitoring as well as in CRM training. Under-communication rather than over-communication is a more likely problem. Assessment of training should address the communication conventions emphasized in training, and these may vary in the degree of flexibility or standardization desired. There may be particular communication gaps or problems targeted in training and these should receive particular focus.

Though communication is important throughout flight, three contexts for communication about monitoring are likely targets for assessment:

1. Is the clarity and completeness of briefings (preflight or inflight) sufficient to ensure that the plan for the scope of the flight being briefed will enable the PM to determine if the flight is being flown as planned? Does the briefing comply with airline procedures and policies?
2. Does the PM notify the PF of discrepancies when noticed rather than delaying until they are fully understood and action is identified? Does the PM communicate not only about specific values that violate expectations but also any broader sense that something unexpected or problematic is going on?

3. Does the PM persist in communicating the existence and nature of the problem to the PF until an appropriate response is provided?

A fourth possible assessment context concerns the PM taking a needed control action or avoiding taking an incorrect control action if communication with the PF fails. This might occur if the PF is extremely preoccupied or subtly incapacitated. Taking control actions may be considered outside the task of monitoring, but certainly falls with the duties of the PM role.

As with other skills, communication can be assessed both for conceptual understanding and for fluent execution. Conceptual understanding can be assessed in 'scripts' where the pilot is prompted with "what do you say now?" as an unfolding event is described. In dynamic settings, assessing training effectiveness may generally be best done at the crew level, though individual assessment can be valuable for some skill. One approach for assessing an individual pilot's skill is use of a confederate co-pilot; this confederate pilot can then intentionally create challenging situations for the trainee co-pilot to notice and respond to. This can be useful if assessing response to relatively rare behaviors that are unlikely to occur spontaneously. However, use of a confederate may be avoided in many cases because the assessment is of the training, not the pilot, and effective communication takes place and can be measured at the crew level. Assessment might also include fluent execution of tasks relying on a standard communications protocol or procedure.

As is generally the case, scenarios need to pose challenges specific to the skill being assessed, need to minimize the ability of the crew to anticipate the challenge, and need to have clear criteria for scoring the key aspects of performance. Measures of communication will overlap with other measures of monitoring since many measures of monitoring rely on verbalization. Measures focusing on communication can use time measures, including time for first mention of an anomaly or unexpected event, time to mention cause or interpretation of the issue, time to mention needed action, and time until needed control action is taken. Not all these communications may occur, and non-occurrence of an "attractive" but incorrect control action may be relevant. Slow uptake of an identified control action that requires persistent communication by the PM is one example where a confederate pilot can be a useful element of program evaluation. Style of communication about important events may be scored as well, particularly if specific guidelines and examples of concise but explicit communications have been provided in training.

It is worth noting the value of conceptual understanding and ability to communicate monitoring process and results, not just fluent execution of routine skill. We believe an important training goal is to make knowledge more explicit. An explicit understanding of monitoring skills provides pilot with common ground, provides shared terms, and shared conventions. This enables situation models that are consistent across crew members and makes communication easier. For example, standard concepts about monitoring roles, shared language such as red/yellow/green status, and shared conventions about how a PM alerts the PF to a problem can all aid communication. This can also make it easier for a pilot to "talk to him/herself" as in self-debriefing, and to reflect on their own performance and how to improve it. Pilots do spontaneously reflect on their performance, commenting "Oh, darn, I should have…" and this meta-cognition is a valuable tool. We suspect that the ability to reflect on and communicate about the current situation will be increasingly important as supervisory control, rather than direct manipulation, becomes a greater part of piloting. Evaluating effectiveness of different training approaches, such as the VVM approach to sharing information, may be increasingly valuable.

4.6. Summary: Performance Measures for Assessing Program Effectiveness

Monitoring is difficult to measure. Nevertheless, a broad array of measures is possible, and no single measure is perfect. Broadly, using a variety of measures of the same activity and using a variety of activities is the most valuable approach for assessing the training effectiveness of a training program.

Measurement can serve the goal of providing a summative assessment of overall training program strength, as currently done in AQP certification. We believe it is also valuable to identify specific training strengths and weaknesses. In the short term, it is useful to know where the untrained vulnerabilities are and hence what situations or activities may carry greater risk. In the longer term, this provides formative guidance on where, and possibly how, training can be improved. In addition, summative measures can be derived by integrating more detailed measures that also support formative evaluation. Thus, assessing component knowledge can provide considerable benefit at little cost.

Measures of integrated performance are important because this is what happens in the operational context. Integrated monitoring skills can be assessed in the simulator as well as in operational flights.

The following Sections 5–9 describe monitoring evaluation at different Levels of Evaluation. Section 5 describes assessment of attitudes, not performance. Later sections give example applications of the basic performance measures described in this section. Section 9 describes additional challenges developing measures at Level 4, operational impact.

5. Level 1: Trainee Opinions and Feedback

5.1. Level 1 Characterization

Level 1 asks participants, usually trainees but possibly instructors as well, to judge the training program. Questions might cover many topics concerning how much a participant liked it, thinks it will be useful, or considers it operationally relevant. More specific feedback about training content or methods can also be collected. This form of evaluation may provide useful information about training strengths and weaknesses and may increase training acceptance. Interpreting these data must be done carefully, as asking for trainee opinion does not measure whether in fact the trainees learned anything.

Pilots might be asked to identify aspects that they thought were particularly good (useful, operationally relevant, clear, helpful to me, etc.) or poor, or to rate each component or activity in training. Specific information can be particularly useful in formative training evaluation. Assessment is usually done at the time and place of training but may also be delayed. Questions may be posed by survey or interview. As an example, the early assessment of NOTECHS training when applied to medicine relied on participant survey (Flin et al., 2007).

Note that Level 1 evaluation generally does not address whether something changed from the training. It is not sensible to ask participants what they thought of the training before they had it to compare with what they thought afterwards. There is a case where change is of interest, but performance is not involved. Surveys might investigate attitudes and whether attitudes were changed by training, which might be considered Level 1. While attitude change may be important in some situations (such as efforts to change safety culture), we expect performance-based evaluation will more typically be of greater interest.

Level 1 evaluation requires decisions about who participates, how data are collected, and what survey or interview instrument to use. Guidance about these topics from the social sciences can aid collection of the most accurate, useful information (e.g., chapters in Reis & Judd, 2009). Visser et al. (2000) provide particularly relevant information on surveys and survey instruments and this is a key source for many of the recommendations here.

5.2. Evaluation Participants

Participants are usually the trainees but getting assessments from trainers can also be valuable. Participants, particularly trainers, can also act as supplemental subject matter experts (SMEs), and some may have informed views about addition, deletion, or correction of content or of delivery method. The most informative survey for trainers may be different than that for trainees. Frequently, all people who are provided the training are included in an evaluation, but if Level 1 assessment is used for a large, fully implemented program it may be necessary to sample participants. Social psychology has developed and applied methods for selecting a representative sample of participants, such as randomized and stratified sampling. For simplicity, we will refer to participants as the pilots.

5.3. Setting

Data can be collected by the individual pilot filling out a survey, through an interview, or by using a combination of the two. For example, each pilot might fill out a survey, followed by a group interview to clarify or extend related information in a briefing-like format with a trainer. The personal interaction in an interview typically provides more flexibility, more motivation, but potentially stronger demand characteristics on the trainee, such as motivation to say what the interviewer would like to hear. Structured interviews follow a series of questions, and the structure is much like carrying out a survey of open questions interactively. Here we do not address the specifically interactive aspects of good interviewing technique. Broadly, the principles for good survey design apply to design of structured interview questions.

Data are usually collected at the end of a training session in the same location as the training. This maximizes response rate and minimizes forgetting about the training format, content, or experience. Interviews or written responses are both feasible in this setting. It is also possible to solicit feedback after a pilot has returned to flying the line for some period of time, which has the potential benefit of trainees noticing whether and how they applied what they learned to their work. Delayed assessment may give more helpful information about perceived usefulness but may have less useful feedback about particular aspects of the training due to memory decrements from the delay. Typically, delayed assessment will produce a lower response rate and would require a survey rather than an interview-based delivery.

5.4. Evaluation Content

Questions may be asked about general characteristics, which are relevant to many types of training programs. Questions might probe perceived usefulness, realism, importance, interest, clarity, or novelty, as well as liking. It is often useful to ask participants about what they thought particularly valuable and least valuable, including comments about possible improvements. Broad questions about liking or usefulness can be helpful predictors of acceptance or level of enthusiasm for the training. In addition, overall reaction to the training may provide initial evidence that it might deliver value.

Part of the content can be more specifically tailored to issues of interest to the people developing or using the training, asking about specific concepts, delivery elements, or activities. Breaking down questions to ask about particular components provides more targeted information than judgments about the training as a whole. In turn, this may provide more useful feedback for improving the training. For example, a survey might ask about clarity of particular topics, such as the model of monitoring, the role of communication in monitoring, or how to identify what to monitor in what situations. A survey might ask about the value of different activities or parts of the training: how useful or interesting was watching the video, doing the debriefing, listing types of monitoring challenges, or providing an explanation?

5.5. Evaluation Instrument

Purpose and format are each important to consider in designing the assessment instrument. The purpose of the evaluation should determine the content of the assessment instrument to ensure that the most valuable information is collected. Some degree of pretesting the assessment instrument is useful, even if done only with a handful of respondents. This is an efficient way of checking length, identifying gaps, and identifying any aspects that are confusing, annoying, or not interpreted as intended. Once a useful survey has been designed, it may be possible to adapt and reuse in future assessments.

5.5.1.Purpose

For the purpose of investigating and possibly shaping global attitude toward the training, the evaluation can ask about its personal relevance or how valuable or interesting it was. Asking for evaluation with clear intent to make use of it may make pilots' experiences more positive by including them in the development process. Assessing their own experience may consolidate the pilots' attitudes (positive or negative) and make that opinion more accessible in talking to others. This can influence how the training is anticipated and perceived by future trainees, and, when positive, may increase the engagement of future trainees.

For the possibly more important purpose of gathering feedback to improve training, the evaluation can ask the pilots questions to diagnose training strengths and weaknesses. Nevertheless, an assessment can also capture information on unexpected topics by including open questions. One helpful approach is to begin with open questions about strengths or weaknesses. This rather directly taps into the pilot's own experience with various training elements. Because these are open-response questions, this allows pilots to draw in additional issues or topics; responses from instructors may be particularly informative here. This may identify what pilots already know, terms they use, new concepts, and activities they thought valuable. If similar topics or elements across pilots are mentioned negatively (as confusing, unimportant, already familiar, boring, etc.), this can provide particularly helpful feedback. While positive information is useful as well, people may be disposed to report positively, and these responses may be less reliable despite efforts minimize this. Questions asking for comparisons to prior training may give good information about how the program was perceived and whether novel aspects were effectively communicated. Engaging pilots to assess what was presented and inviting them to mention what was left out or how to improve can be helpful, though do not expect pilots to design the training!

5.5.2. Format

Several topics about survey form are very helpful to bear in mind when creating (or selecting) a survey for assessing the monitoring training:

1. Open-ended or closed-ended question format. Open-ended format allows the respondent to choose their own words. A close-ended format requires selection of a fixed set of responses, such as a rating or multiple-choice. Open-ended questions in surveys are more informative and as reliable as closed-ended questions. They usually take more of the pilot's time and are also more difficult to score. Thus, if you have resources to code these, you will learn more. Scoring open-ended questions may be prohibitive for very large groups, but they are particularly valuable to get the most information early in training development and implementation. Responses from open-ended questions can identify the most important topics, and these can be addressed in closed-ended responses. Thus, a combination of open- and closed-ended format questions can be used, with particular reliance on open form early in assessment, and where discovering new ideas and unexpected information are most important. Open-ended responses may be more valuable from some pilots than others, e.g., for instructors vs trainees. For example, instructors may have useful feedback about the training to improve as well as validate.
2. Ordering. a) Make the first question engaging, that is, clear, interesting, and not threatening. b) Put neutral questions before questions with "hints." Sometimes more specific questions may directly or indirectly identify what is important, expected, or normal. While information-rich questions like these might be needed, placing these after more-neutral questions means you can get more-spontaneous answers before those influenced by the survey itself. c) Respondents may get tired, bored, or even run out of time. Difficult questions are more likely to be answered thoughtfully if not at the end. There may be no perfect order but bearing these aspects in mind can prevent undue influence from the survey rather than the training it is measuring while helping the respondent provide the most information.
3. Wording. a) Keep questions clear, short, and focused on a single topic. b) Different wording of questions intended to get at the same information can change the framing or perspective of the respondent. Often asking questions about the participant's assessment for him/herself ("useful to me") is the simplest, most direct framing; however, distancing from the self could be helpful if pilots are reluctant to report they needed or benefited from training at all, or about some aspect ("useful for pilots when joining the airline"). Wording also matters for closed-ended questions.
4. Rating and Ranking. Rating scales are typically used for closed-ended questions. Rating scales on 5 (unipolar "how hard") or 7 points (bipolar, "hard-easy") generally offer the best combination of reliability and validity. Response points should have both a number and a verbal label, with the verbal label trying to divide the scale evenly. Avoid wording ratings that are explicitly positive on one end and negative on the other end of the scale, such as selecting agree-disagree or yes-no responses. Instead build in the specific content, even if one part of the scale is more positive. Wording with specific content built in would be rating the prompt "Overall this training is" on a 5-point rating where 5 is labeled "very important" and 1 "not important at all." Wording with a positive-negative scale would be rating the probe "Overall, this training is very important" on a 5-point rating scale where 5 is labeled "strongly agree" and 1 is labeled "strongly disagree." Note that the second rating scale has no specific content but a generic positive-negative scale. If rankings (ranking the importance of different

topics or activities) are desired, asking this directly is most reliable, rather than asking for ratings and computing ranks.

5.6. Summary

Level 1 assessment of training concerns the perceived value and acceptance of the training. This may be particularly valuable when the training is still in development or newly introduced. Level 1 assessment may be used to gather feedback about recommended points of improvement, omissions, priorities, etc. Engagement of the trainees in program development can be helpful both for improving delivery and increasing acceptance. Level 1 data can be particularly valuable if participants include relatively knowledgeable pilots. Surveys and interviews are well-developed study methods.

6. Level 2: Component Skills and Knowledge Underlying Operational Behavior

6.1. Characteristics

From an aviation perspective, the key aspect of Level 2 assessment is focus on component competencies. Further, these components can be measured in settings such as a classroom or pilot's home using relatively simple technology, including paper, video playback and response, laptop-based interactions, or procedure trainer. Computer-based tools can provide an increasing array of interactive assessment methods. Measuring component competences in these more accessible settings and with simpler technologies can provide assessments that are less expensive and may be more diagnostic than assessments in simulated or actual flight. Assessment can be more diagnostic because multiple short items, examples, or tests can be carried out quickly providing a more sensitive measure of the targeted skill(s) or knowledge. Items presented as short videos can provide the virtue of preserving the dynamic character of monitoring but allow the pilot to recognize appropriate or problematic behavior in others, which is a simpler task than generating the behavior themselves. Overall, controlled settings where pace and content can be determined based only on evaluation needs are very helpful for assessment of components. Because program assessment is an iterative process, understanding what components are well or poorly trained is helpful for both summative and formative evaluation purposes.

Note that targeted assessment of components can also be carried out in simulators and even in revenue flight, but diagnostic scenarios are difficult to design (in simulators) or identify (in flight). Briefing and debriefing provide "tutorial-like" settings where assessment of component skills might also be carried out.

Assessment of what component knowledge has been learned is often done as part of the training program itself, possibly as one input to individualized instruction. If assessment of pilots within the program can be coordinated with assessment of the program, this can provide great benefit. A key benefit comes from using pre- and post- tests of individual pilots. Change scores in pilot's component skills and knowledge can then also be used to assess program effectiveness.

Within these assessment environments, there is still a wide range of choices for assessing component skills. For example, knowledge of mental models is needed as an input to a situation model to provide reference points for comparison. Mental models could be assessed quickly but more superficially with well-designed multiple-choice questions, for example, about how autoflight mode impacts performance. A more in-depth measure would be to ask the pilot to type in an explanation of the differences in performance between FLCH, VNAV PTH, and VNAV SPD and in their sources of the target altitude and speed. One challenge to managing complex, open-response questions is difficulty in scoring. A phased data approach can reduce scoring burden: 1) pilot writes and uploads an answer; 2) pilot is given access to a scoring key; 3) pilot annotates their answer, mapping components on the key to components of the answer; 4) pilot uploads the self-scored answer. (Potentially, open-response questions could be converted to conditional multiple-choice questions.) Average pilot performance, with and without training, on different samples of explanation questions would then provide data about program effectiveness.

6.2. Content

As is clear from the high-level characterization of the monitoring competencies described in Section 2, monitoring is cognitive work. This includes not only the knowledge, such as that captured by the mental models, but also the processes. These include processes such as the comparisons for assessing whether the current values in the situation model match reference values from relevant mental models or estimation of descent rate using the three-to-one rule. Even where external actions are a necessary part of the task, the limiting aspects of task performance are cognitive, not motor skills: configuring the displays to provide the best information on approach is not limited by motor skill but by understanding what information is most important and how to display it to best support monitoring and awareness.

Specifying the detailed monitoring content is a difficult but important task that must be carried out by each airline. Conducting this analysis is difficult for at least two reasons. First, many monitoring skills are fairly general, applying in a wide range of conditions and thus do not fit neatly into a hierarchical task decomposition. Second, cognitive skills per se are not directly observable. For reference, effective TEM or CRM does not consist of following well-specified behavioral procedures, but application of good judgment. Similarly, effective monitoring, or situation assessment, depends on understanding and judgment. Despite this, much specific, helpful guidance can be provided about useful strategies in particular situations. Linked training of both the context-specific strategies and more general understanding of the monitoring spiral, or cycle, may be useful; training of both types can be evaluated in program assessment. General skills for carrying out the monitoring spiral are identifying what questions about the situation are important to ask in the current context, how to gather data and assess the evidence collected, and how to identify what actions are or may become needed. In addition, skill with the monitoring spiral facilitates and depends on both communication and task management skills.

Broadly, context-specific monitoring strategies can be tested by setting up the relevant context and testing whether the pilot can recognize or produce the trained strategy. Of course, training is to event types not just a specific event: use of the three-to-one descent rule to estimate feasibility of making a waypoint can be tested in situations that differ in detail from those used in training. Testing more general monitoring skills is more difficult. The basic approach is to provide a situation that was not addressed in training, but one where a pilot with a general understanding can apply that to produce an appropriate response. Clearly, judgment is required in deciding how different the situations used in assessment items should be from those used in training, and also how unexpected or surprising the situation itself might be. Situations posed in assessment questions can differ in how they are presented. For many assessment goals, it is useful to present a relatively rich context (e.g., an image of the flight deck, a description of the problem in some detail) so that the processing needed to answer the question is similar to what will be needed in flight.

6.3. Examples

Following are illustrations of the types and variety of methods that can be used to assess components of monitoring skills. These skills certainly overlap and interact, yet it is feasible to assess where difficulties may lie. These examples center on behavior of autoflight modes. The purpose of these examples is to illustrate that the same topic (autoflight modes) can be assessed for different monitoring competencies (e.g., mental model knowledge, identifying the important monitoring question).

Knowledge about mental models could be tested as fact recognition or in assessing hypothetical situations. For example, the pilot could select the correct multiple-choice answer about what will happen at top of descent if the flight was never cleared to the altitude in the flight plan from a set of plausible choices. Alternatively, the pilot could be shown an image of the flight deck including the current LEGS page, a current clearance, and asked what if any control action should be taken and why. This requires integrated reasoning to set up an appropriate situation model including an understanding of the auto-flight system and to reason, for example, about the intervention needed near T/D to ensure the airplane begins descent. If questions about how the autoflight modes work cannot be answered in relatively simple situations and without time pressure, this suggests inadequate understanding of modes.

It is also important to assess ability to apply an understanding of mode behaviors to guide the process of monitoring. Breaking this down into the three parts of the monitoring cycle provides a method for doing this.

1. *Identifying priority questions* might be tested by providing a short video of a PF selecting a mode in response to an ATC clearance. The pilot being tested might be asked to either generate or to select priority questions. Several candidate questions might be posed and the pilot asked to order them or to mark as high vs low priority. Candidate questions might include: Is this mode what the pilot intended? Is this mode normal in this situation? How will this mode impact flight path? Was this mode stated in the pre-flight briefing? What potential issues does this mode bring?
2. *Gathering and assessing evidence* can be assessed by posing a situation, with an appropriate question to be answered, and asking the pilot about it. Information displayed about modes is relatively limited, and primarily consists of the MCP, the controlling modes at the top of the PFD, and the flight plan in the FMS. Evidence gathering and assessment might concern whether VNAV PTH will meet restrictions at the waypoint after next when the first waypoint might be crossed at the high end of a window, perhaps by including a worst-case computation of the 3-to-1 rule. In other cases, reasoning about making restrictions might be guided by setting up the “green arc” or a Vertical Situation Display, depending on the aircraft.
3. *Assessing action* can be evaluated by providing a question and the displays providing the needed data and asking whether the current control mode will result in meeting the ATC clearance. An alternative question might ask, for a marginal case, what mode will have the greatest ability to meet a clearance. For example, the control law in FLCH might produce the needed steep descent.

Assessment of communication and of task management can begin in classroom or computer-based training (CBT) situations. Indeed, role-playing in a class setting can be an effective training method for communication. Not only is communication an important part of monitoring but talking between the PM and PF provides a natural window into the monitoring process. Assessment of communication can include content, timing, and style.

Assessment of task management fundamentals can also begin in simple and even static conditions. The pilot can be asked to prepare a briefing that plans when tasks will be done and by whom, with a particular goal that the PM's workload does not compromise monitoring. The pilot can be presented with a video or written vignette describing an overload situation and then asked to specify what tasks should be shed, deferred, or reallocated. This might usefully be done from the perspective of both PM and PF.

6.4. Summary

In short, every component of monitoring skill can be productively though incompletely assessed with simple technology and controlled environments, as in classrooms and with CBT interactive activities. Many forms of testing are feasible, as outlined in Section 4, and can provide quick, multiple-item tests of a given capability. Tests may often require a pilot to verbalize information. This might be viewed as an "extra" unrealistic demand. However, practice verbalizing and use of a shared terminology is likely a contributor to effective monitoring. It is possible though that poor design of testing in these simpler, controlled contexts could introduce extraneous test demands not representative in real flight conditions. However, if performance is poor here, it is unlikely to be better in more complex, dynamic, and likely time-pressured environments of simulated and revenue flight.

7. Level 3A: Behavior in Simulations Modeling Operational Work

7.1. Characterization

In the context of aviation, Level 3A means assessment of pilot behavior in a simulator of at least moderately high fidelity. The relevant level and basis of fidelity may vary with the specific skill being evaluated. The availability of flight simulators has transformed both training and performance evaluation, allowing realistic work in a controlled context. Evaluating training in a simulator has important advantages relative to simpler settings and relative to revenue flight (Billman et al., 2019; Mumaw et al., 2019a). It provides a much greater degree of control than in actual flight while still providing high realism. In addition to testing performance in naturalistic flight situations, flight "snippets" can be used to sample performance on many tasks quickly, as in Level 2. The simulation can be stopped to ask a pilot questions. Simulators allow controlled testing in extreme and dangerous conditions and provide a very powerful environment for assessing operational work. Nevertheless, it can be difficult to assess the impact of some operational factors on monitoring, such as state of the airspace or stress and surprise, in this setting.

The simulator environment can be used in several ways. It can be used to model normal flight, perhaps to estimate frequency of occurrence of some behavior difficult to observe in actual flight. It can be used to assess overall performance in particular situations of interest, perhaps because they are dangerous or perhaps because performance in these conditions is judged to be problematic. Finally, they can be used to assess component skills in realistic settings; here "component skill" may be a component of monitoring (as laid out in Section 2), or may be monitoring overall, as a component of piloting. The simulator provides a very powerful environment for assessing monitoring. This last approach is most relevant to evaluating effectiveness of a program for training monitoring.

7.2. The Importance of Scenario Design in Simulation-based Program Assessment

Assessing a component skill requires very careful design of the scenario used in the simulator. To test one aspect of piloting, performance should not be limited by some other aspect. For example, scenarios for testing pilot situation awareness (a result of monitoring) should not include such demanding manual flight skills that performance is very poor for this reason; however, task difficulty can be modulated by increasing demand from other factors, as has been pointed out (Hoermann et al., 2003). Thus, scenarios to test monitoring need to be designed to provide situations that pose monitoring challenges. Performance on the challenge then is likely the element limiting performance on the scenario. If the scenario is handled well, this aspect of monitoring is likely adequate and, if not, that aspect is likely to blame.

Scenarios specify a particular context (phase of flight, weather, etc.). As a result, the monitoring challenges they pose may primarily stem from a particular aspect of monitoring (missing a critical question to ask, poor task management of distractions, not knowing where to find needed information, or failing to communicate a recognized anomaly promptly that requires interrupting the PF).

We consider context specificity first. Any scenario represents a particular situation and the tasks relevant in that situation. Only a small sample of possible situations can be included in evaluation. An important reason for selecting a situation is that: a) the situation seems to pose difficulties based on performance in revenue flights or in training and b) there are particular, known monitoring strategies or knowledge likely to be useful in the situation. Thus, evaluating whether training has been successful in teaching the relevant strategy for these situations can be an important goal. Some illustrative examples are:

- How to anticipate what the autoflight system will do at top of descent (T/D) (or to avoid other ‘automation surprises’).
- What questions to consider, generally, when planning how tasks will be done on approach, or more specifically, assessing the impact of a traffic slow-down ahead and what, if any, actions may reduce the impact.
- The importance of and method for checking that an ATC clearance is feasible.
- Checking the impact of a mode change entered by the PF rather than simply repeating the stated mode.

The second type of specificity relates to the scenario’s specific diagnostic test of a particular component of monitoring. Problems with a diagnostic scenario not only demonstrate problems in the particular situation but may suggest more general problems with that aspect of monitoring. If an inadequate mental model of autoflight modes is diagnosed in one situation it may impact performance elsewhere. If a very passive, superficial investigation of relevant evidence is apparent in one task, it may reflect a more general misunderstanding to the purpose or method for assessing evidence. If pilots in the PM role are late or reluctant to communicate anomalies found in one setting, the same may happen elsewhere. Inadequate understanding of how to check sensor reliability in one situation may signal a broader problem with evidence checking.

In addition to flying extended, realistic scenarios, simulators can be used to test ‘snippets,’ where capturing actions is important. Generally, if action is not being assessed, but only noticing and analyzing the presented information, this can be done with video or animations, which is a much less expensive method for those evaluation goals.

Factors important in scenario design for evaluating an interface have much in common with factors relevant to training-program assessment. Mumaw, Billman, & Feary provide a description of factors for interface evaluation (2019).

7.3. Measures

Well-designed evaluation scenarios for simulators also need relevant measures. The primary type of measure involves pilot performance; that is, observable behaviors. Observable behaviors may be verbal or nonverbal. Verbal behaviors are the questions or comments stated aloud, to self, copilot, or ATC that reveal some aspect of the cognitive activities in monitoring. This might be stating a current variable value, asking a question, or giving a directive. Awareness of the current situation (good evidence gathering) is often revealed verbally. In simulator-based assessment, pilots may be explicitly taught to talk more or "think aloud" more for the purpose of program evaluation. Alternatively, willingness to verbalize appropriate information may be part of the assessment; part of the work of monitoring is communication. Generally, pilots flying the simulator need to fly as they normally would. However, it is possible to modify this to enable better measurement. For example, the simulator might be paused and a pilot asked to specify what will happen without pilot intervention or what the effect of a particular pilot action might be. Specific questions targeting the topics of interest may be asked. There are also standard situation awareness measures that can be employed while the simulation is running (SPAM, Durso & Dattel, 2004) or by stopping and blanking the simulator (SAGAT, Endsley, 1995). These measures may be used or adapted to provide information about the current contents in the situation model. Verbal report in debriefing can be valuable, as well. Because such reports are retrospective, forgetting will occur particularly for information not considered significant and not incorporated into the situation model. Despite the limits of retrospective report, this may provide useful, complementary information (e.g., SART, reviewed in Jones, 2000).

Nonverbal behaviors can indicate the process of monitoring, as well. Scenarios can also be designed so that changing displays or seeking information are indicators of what is being considered. Getting the aircraft on the cleared trajectory and configuration is the ultimate goal of effective monitoring. Thus, taking relevant control actions is also a measure of monitoring effectiveness, preferably by the PF, but by the PM, if needed.

Direct measurement of aircraft behavior can be useful, as well. Sometimes this can be inferred from pilots' actions, but it can also be measured directly using the simulator logs. Many variables are available, and desired aircraft performance can be specified in terms of these variables. Specification could be in terms of whether changes to trajectory are made in the MCP or flight management system (FMS), in the percentage of approaches in which the course is within a specified amount of the cleared course, in whether the aircraft meets each of the waypoint restrictions, in whether flaps are set at the appropriate time and airspeed, etc. Some measures are directly stored in the logs and others might require supplemental data, such as the timing and content of flight management computer (FMC) clearances. Pilot control actions can also be measured using simulator logs rather than the scoring by an observer.

7.4. Summary

Simulators provide a tremendously powerful tool for program evaluation. Pilot performance can be assessed over a very wide range of normal and non-normal situations. A great deal is going on in any simulator flight. Thus, care and precision in the design of scenarios is required so that scenario performance is as diagnostic as possible. In addition, performance needs to be measured in ways that largely preserve the natural activities of piloting while also revealing the strengths and weaknesses of the largely cognitive processes that make up the scope of monitoring as sensemaking.

8. Level 3B: Behavior in Operational Work

8.1. Characterization

In the context of aviation, Level 3B evaluation measures pilot performance in revenue flights. A key constraint here is that measurement must not disrupt performance. Very small interventions during flight may be possible, such as asking a question at a point of low workload. Of course, scenarios cannot be controlled for purposes of evaluation, so there is much less ability to target or control operational situations or the challenges that must be monitored.

Assessment in actual flight is critical to program evaluation. Some aspects of actual flight may not be feasible in a simulator, such as traffic interactions in the airspace, stress, fatigue, or surprise. Naturalistic observation has the great virtue of capturing unexpected, as well as, planned events and is a powerful exploratory method. However, from the perspective of program evaluation, controlled comparison is very valuable. Thus, many more observations of actual flight than simulated flight may be needed to be able to measure changes in performance.

No projects that we are aware of specifically address evaluation of monitoring or of monitoring training. Several behavioral marker systems have been developed for coding CRM or nontechnical skills in operational flight. Some include situational awareness, which is related to monitoring. NOTECHs was a project focused on developing an evaluation method for situation awareness designed to be used across carriers and countries in Europe (Flin et al., 2003; O'Connor et al., 2002) It developed behavioral markers and assessed reliability by comparing the coding of videos by experienced instructor pilots who had been trained in the coding methodology and found satisfactory agreement. The Advanced Crew Resource Management (ACRM) project focused on developing CRM training, but included a behavioral marker rating system (Holt et al., 2001). This consisted of 10-item scoring used in Line Oriented Evaluation to assess effectiveness of their CRM training program. These were used as part of the annual evaluations to compare performance on these items between training groups. Use of the rating system was judged satisfactory. In the United States, LOSA currently provides a mild form of evaluation based on behavioral markers, which particularly focus on threat and error management. In short, development of useful evaluation methods has proved feasible for other nontechnical skills. These methods could be used to assess either individuals or training programs by coding pilot behavior in simulated or actual flight.

8.2. Measures

Many of the measures available for Level 3A can be applied in Level 3B. There are at least three types of data sources from which measures can be derived: observations by a cockpit observer, data recorded by the aircraft, and pilot information in a debriefing, i.e. self-assessment.

Cockpit observation needs to be supported by a structured method for noticing and recording data. This can be expensive to execute if done specifically for program evaluation and may be integrated with observation programs already in place. Specifically, questions targeting the objectives of the monitoring training program might be incorporated into a LOSA evaluation. Monitoring performance is likely harder for observers to score reliably than performance that is more directly linked to specific, highly visible behaviors such as completing a checklist. Training of observers or raters is important, perhaps particularly for more-cognitive skills. There are, nevertheless, several types of behavior—by PM and PF—that can be observed. Some behavior could be coded for every flight. These include anticipatory and predictive actions for expected events. The content of briefings could be scored for elements such as whether the airspeeds or modes planned over descent are mentioned thus giving the PM clear monitoring targets and for whether mode changes are clearly noted (verbally, by pointing, or in the manner trained) by the PM. As another example of routine-use coding, qualitative scoring of the cases where the PM stated what they were checking (the monitoring question) could be noted. While all monitoring questions certainly do not need to be stated, keeping the PF in the loop about PM activity is helpful for building a shared situation model, for managing workload, and for keeping focus on monitoring tasks despite possible disruptions. In addition to routine-use coding, observers could also use special-use codes on events for which particular monitoring strategies had been trained. Monitoring strategies might have been taught that are particularly relevant to: a) evaluating and responding to late ATC changes to flight path, possibly managing through the MCP rather than reprogramming the FMS; b) monitoring when certain types of items are on the minimum equipment list (MEL); or c) early request for relief from a clearance that is not possible.

FOQA data records aircraft parameters in flight. Like simulator logs in Level 3A evaluation, FOQA data preserves information about the aircraft in flight. Since the purpose of monitoring is to ensure correct aircraft states, these are very relevant data; some of the variables capture pilot behavior directly, and others can be used in combination to assess the quality of the resulting flight.

Pilot debriefing can provide retrospective but useful data. Pilots are likely able to accurately report if they noticed a certain event and considered it important or surprising. They clearly cannot report events they missed. Retrospective narratives about strategy may have interesting information and even qualitative information bearing on the training program, such as use of terminology introduced there. If the training included information about debriefing, this could be assessed directly.

8.3. Measurement Context

While the scenario encountered cannot be controlled, as in simulator studies, flights with certain characteristics can be selected, which is particularly useful with large numbers of available flights. If monitoring strategies were taught for particular conditions, flights having or likely to have these conditions could be selected. This would allow selective review of flights where a particular strategy was relevant. Flight selection might be based on observer-entered event codes or on measures derived from FOQA data.

9. Level 4: Operational Impact of the Training Program

9.1. Characteristics

Changes to training are typically motivated by a perceived problem, threat, or inadequacy at the operational level. Ideally, an evaluation should measure the impact of training on the targeted operational threat. For training programs concerning sensemaking and monitoring, the operational goal very broadly is improving safety. Perhaps there were two occurrences of an unsafe behavior that was expected to be very rare (e.g., terrain proximity on approach). Using incidents and accidents themselves as measures will not be very sensitive because they are so rare. It can also be difficult to identify relevant, more frequent events that are precursors of the targeted unsafe event or predictors of the relevant good outcome. Nevertheless, training program design should certainly not be driven by ease of measurement.

There is considerable overlap in Level 3B and Level 4, as assessment of pilot performance may also be considered directly relevant to operational impact. Performance measures for components presented in Section 4 may be helpful in assessing operational impact (for example, conducted during debriefings), though Level 4 assessment likely will focus on integrated performance rather than component assessment.

9.2. Measures

Operational impact level assessment may emphasize integrated performance. Operational level assessments will differ depending on the focus of the training program and the targeted aspects of performance. A great deal of data is collected by airlines, both in terms of the types and numbers of variables and the number of flights measured. Extensive data analysis infrastructure is in place. Thus, a key strategy for assessing integrated performance in revenue flights is both to access data that are already being collected, to capitalize on the data collection process to collect specifically relevant measures, and to introduce minimal modifications to existing processes needed to measure key outcomes.

There are at least three broad sources of operationally relevant measures: creating a measure of a directly relevant pilot behavior but aggregating across a meaningful operational unit, using a measure already in safety reporting, and creating a measure from FOQA or LOSA data. Any ability to integrate data sources is valuable though currently very challenging.

First, the desired operational change might be very close to pilot performance. Pilot performance measures useful for Level 3 may also be helpful here. For example, an operational concern might be reliable pilot compliance with a procedural change. Suppose a new procedure or guidance was issued concerning coverage of pre-descent briefings or criteria for a stabilized landing was narrowed. Compliance with these operational policies can be measured at the level of pilot performance by observers in a simulator or in a LOSA flight. A scorecard for briefing content could be used. For stabilized approaches, airspeed at the time of flap setting is an observable pilot action and is also recorded by the aircraft; precursor actions that facilitate appropriate timing of flap setting might also be recorded, such as time when the checklist was completed. These are measures of individuals' performances, but they can be aggregated and analyzed to provide a group score on this measure. The group might be very selective, such as all flights using a particular STAR, with distance less than some criterion, or with certain items on the MEL; or the group might include an entire fleet.

Second, measures already included in safety reporting may be sufficiently related to the training objective to use as an indicator of the desired outcome. For example, reduction in unstabilized approaches or out of configuration runway accelerations might be one desired outcome from improved monitoring, and this measure is likely already used in safety reporting. Using relevant existing measures, typically from FOQA, is valuable both because the measures are already established as operationally relevant and because there is no additional effort involved to derive, construct, and check a new measure.

Third, new, targeted measures could be derived from FOQA, LOSA, or LOSA-like data, or a combination of the two. FOQA data could be combined with data from other sources. Suppose that reducing cumulative vertical deviation from cleared flight path is a goal of a particular monitoring training program. While a great deal of this information for identifying deviations can be derived from FOQA data, there will be exceptions where either the clearance is not known or where the safer course was in fact to deviate from the clearance. Some of these exceptions may be idiosyncratic (avoiding flight into protected airspace), but some classes of exceptional events might be signaled and broadly screened out (deviations when pilots might plausibly be avoiding icing or thunderstorm conditions). While categories of risky or unsafe aircraft states can often be broadly identified, there are usually exceptions depending on the context. Basically, following standard practice and performing well on a measure (such as minimizing deviation) is usually the safer choice. Nevertheless, there are exceptions to what is the safer choice depending on context. Using context information to screen out likely exceptions can increase the validity of the measure. A distinct, new approach is data mining, which can find differential events, behaviors, or states that discriminate between good and bad outcomes. If such discriminators are identified, human judgment is needed to assess whether these are likely to be relevant program evaluation measures. This data discovery method is relatively new and unvetted but may be a promising path for developing measures applicable to large data sets (for an example, see Stewart et al., 2018).

Capitalizing on the expert observers in LOSA is a valuable resource and might be shaped to the particular needs of assessing monitoring. A caveat is in order, nevertheless: LOSA observers are trained in and LOSA is based in Threat and Error Management. Shifting to assessment of monitoring skills may have pitfalls, and awareness of the different evaluation foci is in order. While explicit scorecards with details of triggers and behaviors are feasible for simulation-based testing, the behaviors for scoring revenue flights will be much more general. Exploring the best resources for operational observation of pilots is an important part of assessment development.

Effective measures need to be linked as closely as possible to both the target of training and to operationally important events. Reports in NASA's Aviation Safety Reporting System (ASRS) and Aviation Safety Action Program (ASAP) are much more useful for suggesting ideas than for evaluation. Unfortunately, ASRS reporting of an event is affected by many factors other than simple occurrence of the event. Events are more likely to be reported when they pose jeopardy to the reporter, when others are also likely to report the event, and when entering a report is not displaced by other demands. Even for the most conscientious reporters, memory of the event will be influenced both by perspective and by passage of time. Despite these limitations, ASRS reports may be helpful in combination with other data sources.

Linking data sources could be very powerful, but is a very challenging goal, due to confidentiality agreements and anonymized data of LOSA and ASRS data. For example, linking ASRS or LOSA data with FOQA data can provide context for understanding FOQA events; uses for this linkage are being explored by some airlines. In turn, this might suggest how different classes of events, such as exception types, might be defined and filtered.

Assessing aviation safety training has some advantages compared to training for other industries and other goals. In some industries, Level 4 operational assessment measures are variables such as a division's costs or company profitability, and these may be rather indirectly related to the focus of training. In calculating fleet level safety performance, an individual flight is the basic unit of analysis. Performance can be summed across flights to measure fleet performance. Further, measures of flight performance are heavily driven by crew performance, though also constrained by context. Performance of the crew and of the pilots making up the crew is the target of safety training. Thus, measures of the process—what the crew did—and the product—what the aircraft did as a result—are both relevant.

9.3. Situations, Scenarios, and Context of Measurement

Data from simulated flights, as well as from revenue flights, can contribute to assessment of operational impact. Simulation testing faces the same challenges in scenario design here as in Level 3B: scenarios need to be designed to be relevant to the issue or factor being tested and to ensure that the performance measure is being limited by that factor. While it may be difficult to know key contextual factors influencing safety in revenue flights, this is usually not a problem in simulated flights. Performance in simulators can provide very accurate measurement of key variables, though from a relatively small number of flights. One important role for simulator assessment is identification of conditions likely to produce particular behaviors (and hence will show sensitivity to measuring them), which can be targeted in revenue flights. Specifying the conditions and behaviors is challenging and is likely to be imperfect. Indeed, measuring monitoring performance in revenue flights will likely have confounds. However, assessment in revenue flight will likely have the advantage that a large number of cases is available as well as directly assessing the performance of in the setting of greatest practical importance. Further, revenue flights can provide base rates (frequencies of occurrence) for targeted behaviors and also for the situations in which the behaviors should occur or be avoided.

In revenue flight, it is not possible to control the scenario, of course, but there is an important analog: selecting what flights should be included in an aggregate measure after the fact. All the flights in a particular time period might be relevant, or perhaps only a small subset. Some aspects of monitoring training may be very general, relevant to all flights, so including all flights in an aggregate measure may be the right choice. Other aspects may target monitoring in particular situations or triggers, such as icing conditions or heavy traffic in merging STARs. In these cases, looking at change in performance on all flights would be a very insensitive measure, and identifying flights in the targeted conditions would provide a much more informative measure of training effectiveness. Thus, conceptualizing what types of flights are relevant and specifying how the flights of that type can be identified is important. Selecting what flights are relevant is an important part of measuring training impact at an operational level.

Monitoring is a step removed from control actions that produce change in aircraft behavior. Measuring monitoring through aircraft behavior is necessarily indirect. Therefore, it may be particularly important to identify the particular flight phase in the particular type of flight where a particular monitoring skill is likely to have an effect. Broadly, this will be situations where the autoflight system flying the initial flight plan would not produce a safe outcome, that is flight phases where the pilots needed to take action. Within this, more-specific conditions likely need to be further identified. Specific monitoring skills taught might prioritize monitoring questions to review at top of descent; when ATC gives a clearance to take the aircraft off a STAR or when held low in cruise; how best to configure displays for arrival; or verifying PF actions when leaving or returning to a STAR. Very specific strategies might also be useful such as monitoring to avoid descent below class B airspace on a specific approach. Scoring is then based on: a) aircraft flight variables relevant to the specific situation type and b) criteria for identifying the flight phase to be scored.

Evaluation of operational impact of monitoring would take place in a data-rich environment, where multiple data types are already routinely collected. This includes training records, LOSA evaluations, and FOQA. Leveraging the data and the data collection process is a powerful resource. However, each of these data collection and analysis processes serves multiple functions, and it is not clear how best to add in evaluation of monitoring training. Integration across data sources is a particular challenge to using existing data sources to evaluate effectiveness of a monitoring training program.

Collecting data from many flights across a fleet has a cost, and the incremental cost of scaling up will depend on the measures already collected. For FOQA-based measures, the primary costs are development of the measures and of the flight-selection criteria and management of the resulting data. Thus FOQA-based measurements are not dramatically affected by number of flights included. Other observer-based data might be needed to provide critical data. Cost can be reduced if data collection can be integrated with other activities, such as activities that are part of the training program itself or LOSA data collection. Coordination of program evaluation with the training program is valuable for every aspect of evaluation, and particularly so here. Note that for established training programs, AQP requires coordination with safety management.

10. Design Principles for Evaluation Levels 2–4: Assessing Training Program Impact

10.1. The Importance of Study Design and Control of Unrelated Factors

Levels of Evaluation 2 through 4 ask whether training produced measurable performance differences and whether those differences can be attributed to the training program. We have been addressing the choice and execution of relevant performance measures, which is a very important aspect of training evaluation. A second important aspect of training evaluation is whether performance improvements can be attributed to, that is, caused by, the training program and is not the result of other factors. **We assume that measuring program effectiveness is the primary goal of program evaluation.** Thus, methods that do this are important. There are design methods for data collection (and analysis) that allow conclusions about cause. Further, in many cases these design methods are practical within the context of program evaluation. We begin with a description of designs for data collection that provide the most useful data.

Good evaluation design, in fact, can do a good job of separating the factor of interest—the training intervention—from other potential contributors to performance improvement. What are these other factors? They include:

- level of initial, pre-training performance
- transient memory of the pre-test
- learning from the pre-test, not the training
- influence of changing operational events or conditions between pre- and post-test on performance
- learning from experiences between pre- and post-test that changes performance
- large individual differences among trainees including level and type of flight experience, motivation for learning, opportunities to learn from co-pilots, and experiences and education before becoming a pilot

We review the possible impact of these factors and then describe how study design can eliminate their effects (practically speaking).

10.2. Diagnostic Study Designs: Did the Training Have an Effect?

We draw on data collection designs that originated in laboratory science and were intended to separate cause or influence by one factor from other possible influences. These powerful designs were developed for very different situations yet they are directly applicable to the problem we are faced with: does training cause improvement to performance?

The designs we will consider are called within-subject, between-subject, and mixed designs. Within-subject designs compare performances by the same pilot (such as before vs after training); between-subject designs compare performances by different pilots (such as a trained and an untrained group); and mixed design uses both types of comparisons. We will sort through how each type of design can help, starting with the simplest and building up to the more complicated but more powerful mixed designs. Each of these can be a practical choice in the airline context, and we provide some illustrations.

First, good performance after training is no guarantee of training effectiveness. The trainees may have been equally competent prior to training. While intuition may suggest this was not the case, without a measured change, no data-based claims of benefit from the program can be made. Data to compare performance with and without training are clearly required. Second, improved performance by the same individual after training does not ensure that the training produced the improvement. Designs that compare performance of the same individual on different tests or at different times are examples of within-subject designs.

To show improvement by an individual, one needs a pre-test (before training) and a post-test (after training). The two test versions (pre- and post-) need to test the same material and be of comparable difficulty. Even without training, simply doing the test a second time might produce improvement. The impact of this type of specific memory will be greatest over short intervals, before details are forgotten. Indeed, immediate improvement may be based on short-term recognition of the type of tasks just used in testing and might occur without any of the training program at all. While learning from the tests themselves might possibly be useful, this: a) says nothing about the value of the training intervention per se, and b) is unlikely to be applied to performance later or in different conditions. The first step around this difficulty is to separate the pre- and post-test in time, either moving the pre-test to be much before training or the post-test much later. This means that specific memory for the pre-test is very likely to fade before the post-test. Moving the post-test quite a bit after training provides an additional improvement to assessment. If improvement is found, it means that the difference after training is still found over a longer retention interval, which is a useful finding. While delayed post-test assessment does not eliminate the possibility of effects from the test itself, it suggests useful learning resulted from training, test, or both; indeed, the testing may be considered part of training. On the other hand, if there is a long interval between pre- and post-test, other intervening factors might influence performance. Change in operational conditions, reported accidents, or other training may intervene and influence performance of a group.

While there are challenges to assessing program effectiveness by comparing the same individuals before and after training, within-subject designs have a great strength. Variation across individuals can be very large and testing the same individual twice eliminates much of this individual variation. If a great deal of variability comes from factors other than the factor of interest—effectiveness of the training program—it makes it difficult to detect meaningful change on that factor.

A between-subjects design eliminates any effect of repeated testing by comparing two groups of individuals, as each person is tested only once. This between-subjects alternative compares a group of individuals who had the training to a control group of individuals who did not. While this removes the possible effect of repeated testing, it increases the influence of individual differences. This means that the people in the two groups should be similar in other relevant respects and that the groups will need to be fairly large to detect a change.

A mixed design can combine the strength of within- and between-subjects designs. Within-subject comparisons reduce the effect of individual differences. Between-subject comparisons remove effects of retesting. A mixed design is somewhat more complicated to execute but can provide a very powerful method of program evaluation.

10.3. Examples of Useful "Mixed Designs"

The mixed-design examples described here are useful in Levels 2–4. All these designs involve two groups of pilots. Two mixed designs are shown in Tables 1 and 2; for easy reference we label the cells A, B, C, and D. These designs might be carried out at large or small scale. For example, 50 pilots might be in each group and the training intervention with pre- and post-tests (as needed) might be scheduled for three hours on the same day as First Look in an AQP assessment. The assessment, for example, might be evaluating a training unit on managing monitoring challenges on approach that presents concepts with exercises critiquing a series of short videos showing strong and weak monitoring behaviors for the trainee to assess, with feedback provided. Pre- and post-test might be flying several short sequences in the simulator.

Table 1 shows a simple mixed design in which each participant is evaluated in a single session. Pilots in Group 1 are compared on pre- vs post-test performance: Did these pilots improve (a within-subject comparison)? Group 2 does not get the training. Group 1 scores on the post-test are also compared to the stand-alone test for Group 2 (a between-subjects comparison): Do people who got the training do better than people who did not get the training (and also did not do the pre-test)? This design has the practical advantage that each pilot only needs to come to one session. Not all the pilots get the training, which may be perfectly acceptable for a new, experimental training program. Pilots in Cell B may be given a control activity of similar length to the training to separate the administration of the two versions of the test.

Table 1. Data Collection: Mixed Within- and Between-Subjects Comparisons: Single Session

	Session 1 (Month 1)
Group 1	Cell A Version A (or B), as pre-test New Training. Version B (or A), as post-test
Group 2	Cell B Version B (or A), stand-alone test Version A (or B), stand-alone test

Table 2 shows a more complicated and more powerful mixed design, involving two sessions. Ideally, the two sessions are run some significant time apart (perhaps weeks or months) to measure retention over an operationally meaningful interval. Sessions might be scheduled with recurrent training. The main change in the Table 2 design from the Table 1 design is bringing participants back for a second session. This gives valuable information, and it allows pilots in both groups to get the training; however, it may add scheduling difficulty. Overall, fewer pilots need be included in each group, because both groups are eventually trained, and both are tested before and after training.

As in the design in Table 1, a) Group 1 gets the training in Session 1, with pre and post-test in that session and b) Group 2 does not get the training in Session 1, which provides untrained scores (on a "pretest") that can be compared to Group 1 post-test for a between-subjects comparison.

However, in Table 2, Group 1 is then tested again in Session 2 in a second, delayed post-test; this can be compared to their Session 1 post-test, which measures within-subject retention and forgetting over a longer interval. Thus, just adding Cell C (without D) adds valuable, distinctive information about retention. In Cell D, Group 2 is now provided the new training, and they then get a Post-test right after training; note this is a long time after their pre-test. Thus, performance on the post-test is unlikely to be much influenced by having taken the pre-test six months earlier. Cell D provides a second, within-subject comparison, now for Group 2.

Table 2. Data Collection: Mixed Within- and Between-Subjects Comparisons: Two Sessions

	Session 1 (Month 1)	*Session 2 (Month 6)*
Group 1	Cell A Version A (or B), as pre-test New Training Version B (or A), as post-test	Cell C Version C, as delayed-post-test
Group 2	Cell B Version B (or A), as pre-test (before Session 2)	Cell D New Training Version C, as post-test

Test versions are counterbalanced, so each version is used in every role, e.g., Version A is sometimes pre- and sometimes post-test. For simplicity, these designs assume comparisons are made with and without the training to be evaluated rather than comparing one type of training intervention to another type of training; the logic is very similar for both cases.

10.4. Mixed Designs for Level 3

Table 3 illustrates how the mixed design for the evaluation shown in Table 2 could be implemented in the context of ongoing recurrent training. For this design, all pilots get the new training but in two waves.

Table 3. Two-Session, Mixed-Design Integrated with Training

	1st Wave-March Recurrent	*2nd Wave-August Recurrent*
Group 1	#1-Monitoring pre-test New Monitoring training #2- Monitoring post-test	#4- Monitoring delayed-post-test Training open/none required
Group 2	Old training #3- Monitoring post-test	#5- Monitoring pre-test New Monitoring training #6- Monitoring post-test

This design compares Old and New training. The design assumes that the Federal Aviation Administration (FAA) has not mandated the New training, but that an airline, probably in cooperation with the FAA, or the FAA itself, is assessing possible training improvements. This design enables several informative comparisons that provide a great deal of information about the effectiveness of the training intervention. Each group is tested three times, once on each of three versions of the test. Order of testing is counterbalanced.

Within individuals, the performance of Group 1 on the March pre-test (#1) can be compared to the March post-test (#2), and each of these tests can be compared to the delayed-post-test (#4) at the start of the August Session. If training is effective, the Group 1 March post-test (#2) would be better than the pre-test (#1). If the training is retained, the Group 1 August delayed-post-test (#4) would be better than the Group 1 March pre-test as well, though likely not as good as the March post-test (#2), because that occurred immediately after training. Similarly, the performance of Group 2 on the August pre-test (#5) can be compared to their August post-test #6).

Between individuals, the performance on the March post-test can be compared between Group 1 trained (#1) and Group 2 old training (#3). If Group 1 learned from training, they (#2) should perform better than Group 2 (#3). In addition, Group 1's test in August (#4) can be compared to Group 2's pre-test in August, to see if any advantage of the new vs old training is maintained after a delay.

Table 4 shows a similar design, with an additional goal to provide pilots with the training quickly, while the evaluation is ongoing. This has the cost that the comparison groups are not being measured at the same times, and different events will have happened between the tests being compared. Testing at different times may be impacted by factors such as change in STAR design, other changes in airline policies, change in nature or scheduling of revenue flights, etc.

Table 4. Mixed-Design. Rapid Adoption a Goal as well as Evaluation

	1st Wave-March Recurrent	*2nd Wave-May Recurrent*	*1st Wave-August Recurrent*	*2nd Wave-Oct. Recurrent*
Group 1 (begins first)	Old training. Monitoring "post-test" (Version A)		Monitoring pre-test (Version B) New monitoring training. Monitoring post-test (Version C)	
Group 2		Monitoring pre-test (Version C) New monitoring training. Monitoring post-test (Version A)		Monitoring pre-test. (Version B) Old (or New) training

The goal is to assess, on groups of pilots, whether a new training program has the desired outcome. Participants are pilots who are scheduled for recurrent training every six months. Groups 1 and 2 are groups of similar pilots each scheduled for training in spring (March and May in the table example). Group 1 starts first and does normal recurrent training (Old) in March followed by a Monitoring Test; this could even be started while the new training program is getting a final polishing. Group 2 does the new extended training in May. This begins with a modified First Look, which includes the scenarios designed to pose monitoring challenges, as well as a variety of normal events. Hopefully, the assessment required in the "standard" First Look might be negotiated in some manner to reduce the time it requires. The training for Group 2 in the First Look includes the *New* monitoring training. As with the design in Table 2, versions need to be counter-balanced, so versions A, B, and C are used equally in each cell.

Fitting the evaluation within the scheduling of recurrent training would have several advantages. First, using the First Look portion of AQP training would provide a no-jeopardy setting with which pilots are already familiar. This is important since the goal is to measure performance of the training program, not the pilots. Second, tapping into the existing training cycle removes the burden of scheduling, and in particular, scheduling a second session at a fixed interval from the first. Third, because a large number of pilots are trained every month, a large sample of trainees can be included. Fourth, cohorts trained at close but different months can be checked for any large differences, such as falling right before and after a recent large onboarding or merger (or occurrence of a global crisis). Finally, existing equipment, data management, and even staffing can be used. The key challenge is negotiating how this can be fit into the First Look sessions and getting buy-in from all involved parties, including airline training departments and pilot unions. A much less satisfactory option would be scheduling the evaluation sessions immediately after the First Look session.

10.5. Design Challenges and Examples for Level 4 Evaluation

Level 4 evaluations have practical challenges to implementing the most informative designs. We believe it is possible, however, to conduct informative Level 4 evaluations with modest additional cost above the development and delivery of the training itself. We think this is true even if measures include observer-based data, not just FOQA data. In particular, we believe the additional costs can consist primarily in data management, scheduling, and analysis rather than in requiring significant additional simulator or observer time. Doing this most efficiently may depend on the flexibility allowed in AQP plans.

Any of the designs described for Level 3 are useful here. An important added bonus is that the testing is not an additional, special activity, but is the normal line operation. This requires good in-flight measures, of course, but means that the measurement, whether observer-based or from data-logging, are part of normal work and less likely to impact performance itself. The examples sketched focus on recurrent training; design details of evaluation plans would be different, but the same evaluation logic could apply for certification or instructor training. We would anticipate that development runs using a small number of pilots would be required to ensure training could be effectively delivered and predictably scheduled, and that this would be needed prior to a broader evaluation.

Airlines have the tremendous advantage of volume: many pilots are necessarily trained every month. This data stream provides a potential resource for evaluating training effectiveness.

10.6. Challenges and Opportunities

AQP training programs provide both structure and flexibility for assessment objectives at particular testing cycles. This structure could be leveraged to assess effectiveness of new training programs at the operational level. This may provide the best opportunity for diagnostic assessment at the operational level of monitoring training. This framework appears to provide the flexibility needed for an informative assessment of training that targets skills or activities not defined as terminal proficiency objectives.

It is worth noting that the role of testing is complicated. For the purpose of evaluating a training program, it is desirable for the test to have no impact on the trainees at all. That is, taking a test (without the training) would ideally not affect performance on later post-tests. However, from the perspective of training pilots, it is desirable for every interaction to increase learning, and testing, particularly briefing and debriefing, can be a powerful learning mechanism. These competing goals are yet another reason why it may be worthwhile using a more complex design where the effect of testing can at least be measured if not removed.

11. Method Summary: Useful Combinations of Context, Materials, and Tasks, with Illustrative Examples

In this report we have described a large number of ways in which the effectiveness of a training program can be evaluated. Here we summarize when different approaches are useful by using three dimensions. Our concern is evaluation of training programs. However, the resources available for training evaluation, as well as specific evaluation objectives, are likely strongly related to the resources available in delivery of training, so these may be closely related.

At a high level, we believe a great deal of very useful information can be gathered in less expensive contexts through strategic design of assessment activities and materials. A variety of assessment activities can be provided on computers, in classrooms, and with less expensive simulators. We think these methods have great potential for development and use in evaluating program effectiveness, both the teaching of components and teaching the integration of components into broader capability. With respect to activities, assessment activities are more usefully spent on applying trained skills, knowledge, and attitudes than on simple memory retrieval tasks; further many of these more active, interactive tasks can still be done in the simpler, less expensive contexts. These evaluations can inform and focus the more expensive evaluation of performance in realistic contexts.

We use three dimensions to organize possible types of evaluation and which combinations are most useful. These dimensions are Context and Technologies, Materials, and Tasks.

11.1. Context and Technologies

We organized much of this report in terms of Kirkpatrick's levels. These were primarily focused on the evaluation goal, but Kirkpatrick linked these to the context and technology available for collecting evaluation data. For many practical purposes, it is the available context and technology that are most directly important for designing evaluation. Thus, Context and Technology is an important dimension. For performance-based evaluation, we consider four categories of Context and Technologies associated with the levels described in this report as Level 2, Level 3A, Level 3B, and Level 4. Table 5 summarizes the levels on this dimension.

We divide Context and Technologies for program evaluation into four groups as follows:

1. *Classrooms and Computers* is a broad and heterogenous group (associated with Level 2) that generally does not require expensive equipment or high-fidelity simulators or situations. This includes: a) classroom environments and the social interaction that it provides; b) traditional Computer-Based Instruction; and c) use of procedure or part-task trainers and low-fidelity simulators that may be available to individual pilots. Development of good evaluation materials may be somewhat expensive but the data collection is relatively inexpensive.

2. *Simulators* refers to program evaluation methods that can be conducted in simulators that are high-fidelity with respect to the training targets. For training monitoring, accurate behavior of autoflight mode transitions might be very important, but precise flight dynamics or full-motion simulation might not be needed.
3. *Revenue Flight* refers to measuring performance in flying the line. Improved piloting of revenue flight is the typical goal of training. Observation of pilots in the cockpit is usually more difficult, more expensive, and provides fewer challenging situations than observation in a simulator.
4. *Fleet-Level Operation*. Impact of a training program at the organizational level can be measured in terms of fleet-level, or higher, operation metrics. Direct measures of safety accidents and incidents—are not very sensitive because rates are so low. Further, outcomes are affected by many factors besides training effectiveness. Indeed, we have not heard of high demand for assessment of training effectiveness at this level, likely because of widespread appreciation of some difficulties. However, impact on established safety indices provides a potential evaluation approach, as outlined in Section 9, and is an approach which might not add substantial cost beyond collection of metrics already in place for other purposes, such as safety management.

Table 5. Context and Technologies Used in Evaluation

Context	*Technology Description*	*Associated Level of Evaluation-Goal*
Classrooms and Computers	Can include much more than traditional delivery such as reading power point slides or listening to lectures.	Level 2
Realistic – Simulators	High-fidelity simulators and targets flight scenarios.	Level 3A
Revenue Flights	Human observation (e.g., LOSA) and recordings (e.g., FOQA) from the cockpit.	Level 3B
Fleet-Level Operations	Aircraft data, training time, SMS safety indicators. Data collection and aggregation technologies.	Level 4

11.2. Materials

Evaluation materials are the questions, problems, and situations used in evaluation tasks or activities. We divide types of materials into three groups that differ in simplicity and in how closely they resemble actual flight conditions (Table 6). We label these Static, Dynamic, and Controllable types of materials. Static materials present events used in tasks and cases only as static descriptions or diagrams. Dynamic materials include video, animations, or recordings. Controllable materials allow the person to affect the course of events, as in a simulation. These distinctions are particularly important for evaluations that include case and event-oriented problems and assessment activities.

Table 6. Materials Used in Evaluation

	Description	*Particularly Helpful for Evaluating:*
Static	Text and diagrams. User can control timing.	Measuring basis for reflection, understanding, and analysis.
Dynamic	Animation, video, recording.	Monitoring transient information, time to notice, interpret, and verbalize.
Controllable	Pilot has control of actions affecting events.	Whole cycle of perception, understanding, and acting.

11.3. Tasks

Evaluation requires some task or activity using the materials where performance on the task or activity produces the data. Pilot performance can be assessed in many tasks, which we group as follows:

- memory (recall and recognition)
- understanding (predicting, explaining)
- assessing physical situation (aircraft config, modes, weather, etc.)
- assessing social/behavioral situation (communication, task overload, etc.)
- action (control actions, information gathering actions, seeking help, etc.)

Performance at the level of operational units can be assessed by measuring ongoing fleet activities. Safety indicators can capture performance on fleet activities. “Activities” at the fleet level might be fleet-level adoption of a new briefing practice, or improvement on a target safety indicator used by the safety management group. See Table 7.

Table 7. Tasks Used in Evaluation

	Task or Activity	*Description*	*Particularly Helpful for Evaluating:*
Pilot Performance	Memory: recall and recognition	Retrieval of information presented in training, through recall (free response, short answer, reproduce diagram) or recognition (multiple choice, forms of true/false).	Mental models, specific strategies or procedures, examples and what they illustrated.
	Understand: explain predict	Explaining what produced a current situation; predicting what will happen, with or without different impacts.	Application of facts to new situations. Reasoning with a situation model and the monitoring cycle.
	Assess situation	Using understanding to produce an assessment of the aircraft and environment state.	Reasoning and decisions about situations, particularly unfamiliar ones.
	Assess behavior	Using understanding to identify problems in the behavior or needs of the other pilot.	Communication, teamwork.
	Act: configure, control	Taking actions needed, including communication, control actions, information gathering.	Third part of the monitoring cycle. Benefits from scenarios where the situation requires a specific action. Task management. Integrated application.
Operational Performance	Operational measures, aggregated or indirect.	Direct and indirect effects of training on measures of safety. Monitoring training can be designed to focus on examples and situations in which problems are suspected.	How well learning objectives and outcomes aligned with the identified operational safety issues, and their measurement.

11.4. A Map of High-value Evaluation Components

The targeted improvements from a training program should, of course, drive design of its evaluation. Areas judged particularly important to improve, areas trained by novel methods, or particularly novel content may be the focus of evaluation. In addition, practical factors such as where data already exist and costs of data collection and analysis are important. Thus, choices of what to emphasize will depend on the particular situation.

Despite variability in needs there are several evaluation tactics likely to be very helpful. Most broadly, a great deal of useful evaluation can be gathered with less expensive and lower technology resources. However, this depends on using evaluation activities and materials that are well-designed for testing skills and knowledge important to monitoring.

Table 8 provides a summary of method choices for evaluation. Cells in the table show the particular combinations of context, type of materials, and tasks or activities in assessment methods. Cells showing combinations likely to be informative and efficient are green and labeled “Good.” This guidance is very broad and some assessment methods from red or gray cells may also be useful. Several points from this table are described before Tables 9–11, which fills in examples.

Table 8. Overview of Good Combinations of Assessment Context, Materials, and Tasks

		Type of Assessment Context												
		Level 2: Classrooms & computers			Level 3A: Lvl 4 simulator			Level 3B- Line flight			Level 4- Ops			
Assessment Task		Static	Dynamic	Control	Static	Dynamic	Control	Static	Dynamic	Control	Static	Dynamic	Control	
Pilot Performance	Memory: recall & recognition	useful	useful	🟥	▨	▨	🟥	▨	▨	🟥	▨	▨	🟥	<---------- deeper, more flexible skills & knowledge
	Understand: explain predict	Good	Good	🟥	▨	▨	useful	▨	▨	🟥	▨	▨	🟥	
	Assess situation	Good	Good	useful	useful	useful	Good	▨	▨	🟥	▨	▨	🟥	
	Assess behavior	Good	Good	useful	useful	useful	Good	▨	▨	🟥	▨	▨	🟥	
•	Act: communcate, configure, control	🟥	🟥	useful	▨	▨	Good	▨	▨	useful	▨	▨	useful	
Ops	Operational measures	▨	▨	▨	▨	▨	▨	▨	▨	useful	▨	▨	Good	
				<---------						---------->				
		more diagnostic of training gaps						more directly relevant to operational safety						

Colors provide a broad index of general usefulness (informativeness and efficiency):

Dark Green = GOOD likely high value.
Light Green = USEFUL also likely to be useful.
Red = unlikely to be a good choice.
Grey = usually not feasible or sensible use of the evaluation environment.
Grey and red categories may be useful for special cases.

Table 8 illustrates the following points:

- Activities that depend on deeper, more flexible skills and knowledge are generally more valuable.
- Very useful, highly diagnostic information can be gathered using low cost settings and technology. To do this, the evaluation can use demanding tasks and the most interactive technologies available in classroom and on an individual pilot's computer or low-fidelity simulator or trainer. Many of these tasks can often be run quickly and inexpensively. Tasks can be designed to assess specific, targeted skills and thus provide highly diagnostic information. These are very useful contexts for evaluating a training program.
- Using the "lowest technology" sufficient to evaluate the aspect of interest can provide effective and inexpensive evaluation data.
- It is possible to use both static and dynamic materials in traditional classrooms and with individual computers. Control tasks are only feasible if some form of simulation is used. However, simulation can often be introduced in these low-cost environments

through the use of role-playing simulation in classrooms and low-fidelity part-task trainers or desktop simulators.

- Broadly, if the evaluation is making use of accessing the more-realistic environments that provide control, it is usually sensible to use tasks where pilots are making use of control. Note that here we mean not only flight control of an airplane trajectory, but also control of the interface and display configuration.
- In high-fidelity (e.g., Level D) simulators and in line flight, the natural assessment relies on the dynamic control these environments provide. The light green cells in the Simulator columns illustrate the possibility of introducing ancillary materials that are just static or dynamic. For example, tasks when the simulator is paused or tasks included in debriefing might ask the pilot to assess (and possibly review on video) what they had just done.
- As is widely recognized, operational-level evaluation of training is difficult because so many variables affect operational outcomes and because monitoring in particular has indirect effects. Nevertheless, standard operational safety metrics can be used (e.g., unstable approaches, hard landings). In addition, aggregating performance from line flights may contribute operational level measures, particularly where change in fleet operation (e.g., how briefings are done) is a training goal.

Tables 9–12 provide illustrative examples. The purpose is to illustrate what is meant by particular combinations of Context and Technology, Materials, and Tasks and thus unpack and illustrate the suggestions shown in Table 8. Each Context and Technology level is shown in a separate table. For the Line Flight and the Operational evaluation levels, using simple materials or tasks may not be the best or even sensible use in these contexts.

Table 9. Context and Technology Level 2: Classrooms and Computers

Static: Pictures, Descriptions	*Dynamic: Videos, Animations*	*Control: Part-Task Trainer, Classroom Role-Playing*
Memory: Recall and Recognition		
When monitoring flight path, what indications are needed to assess making the next waypoint? Circle on the image.	Replay 2 videos from training; which uses the more helpful display configuration for this situation?	(lower value)
Draw and label the Monitoring cycle diagram.	Replay 2 videos from training; which demonstrated good use of the Monitoring Cycle model and which did not?	
Multiple choice questions about how autoflight modes transition in different conditions.		
Understand: Explain, Predict		
Provide narrative description, ask for explanation of what likely leads to this situation. Explain how traffic management is likely to change at airfield X as traffic levels change.	Video snippet (PF enters new altitude but does not change the autopilot or flight director mode): As PM, report what you are noticing.	Pause part-task trainer, report values of key variables; predict what will happen next.
Provide sequence of flight deck displays, ask what info PM should provide to PF.	Students verbally state how the crew used (or did not use) the Monitoring Cycle model after watching an example video	
Assess Situation		
You are CA with Jr FO as PM. You've flown into this airport frequently, what pointers/info might help the FO monitor.	Video snippet: You are CA with Jr FO as PM you've flown into this airport frequently, what pointers/info might help the FO monitor.	Pause part-task trainer: identify priority aspects of situation that should be the current focus of monitoring.
Provide sequence of flight deck displays, ask what info PM should provide to PF.	Video snippet. 1-What do you need to know most in this situation? 2-How do you configure displays? 3-What comparisons and expectations should you check?	Classroom role-playing: PM has relevant information the PF does not. Role play providing and requesting needed information.
Assess Behavior (of Others or Self)		
(lower value)	Video snippet: what did PM do well and what should be different.	Self-debrief after part-task activity,
Act: Configure, Control		
(lower value)	Animation or video show evolving situation. Role-play communication between PM and PF.	Tasks in part-task trainer if timing not focus.
Given image of displays and description of situation, what should be communicated and what actions taken.	Given video snippet, what should be communicated and what actions taken; pilot narration at end.	Record and score pilot performance on scenario snippets.
Operation Measure, Aggregated or Indirect		

Table 10. Context and Technology Level 3A: Using LV4 Class Simulator

Static: Pictures, Descriptions	*Dynamic: Videos, Animations*	*Pilot/Crew Controls Simulator*
Memory: Recall and Recognition		
Understand: Explain, Predict		
		Following a startling event, describe different, relevant expectations (new Mental Model) to 'reset' the Situation Model. Utilize carefully designed events, where correct action will not be taken unless situation is accurately assessed (e.g., unexpected loss of visibility/display/system); ask for assessment.
Assess Situation		
	Video review.	Ask pilot to explain what they are thinking, as if making a training video. Ask pilot to coach or explain to junior FO
Assess Behavior (of Others or Self)		
	Video review.	Crew self-debrief from memory. Does pilot communicate issues around key events built into scenario
Act: Configure, Control		
		Specific: does pilot (re) configure displays to maximize info use? Broad: Scenarios designed to reveal vulnerabilities. Pilot's communication and control actions are assessed.
Operational Measures, Aggregated or Indirect		

Table 11. Context and Technology Level 3B Line Flight

Static: Pictures, Descriptions	*Dynamic: Videos, Animations*	*Pilot/Crew Controls*
Memory: Recall and Recognition		
Understand: Explain, Predict		
Assess Situation		
Assess Behavior (of Others or Self)		
Act: Configure, Control		
		Observation of flight deck.
Operational Measures, Aggregated or Indirect		

Table 12. Context and Technology Level 4: Operational/Fleet Impact

Static: Pictures, Descriptions	*Dynamic: Videos, Animations*	*Pilot Controls*
Memory: Recall and Recognition		
Understand: Explain, Predict		
Assess Situation		
Assess Behavior (of Others or Self)		
Act: Configure, Control		
		Observation of flight deck.
Operational Measures, Aggregated or Indirect		
		Safety metrics.

12. Conclusions

12.1. Our Findings

Any evaluation of a training program must begin with a characterization, implicit or explicit, of what is to be trained. Our qualitative model simplifies the many details of monitoring to provide an organizing framework for training. The model characterizes monitoring as dynamic sensemaking and further identifies the following component competencies:

1. Operational knowledge and experience, particularly the mental models that provide the reference and expected values for comparison to the current situation.
2. The processes of building and maintaining a dynamic situation model based on relevant mental models.
3. The cycle of updating the situation model by posing relevant questions about the operational environment, by gathering and assessing evidence to answer those monitoring questions, and then by assessing the implications for action.
4. Communication and task management activities that enable effective monitoring in a dynamic operational setting.

Thus, any evaluation of a program for training monitoring needs to address these skills and knowledge.

We have emphasized the benefits of including multiple assessment methods and contexts to address a range of evaluation goals. We pointed out how Kirkpatrick's evaluation levels can be adapted for commercial aviation. In particular, we believe it is valuable to assess both individual competencies and integrated, or overall, monitoring performance. Component skills and knowledge can often be usefully investigated by performance measures in simpler, less-expensive settings, such as classrooms or remote learning, as well as in simulators. Integrated performance benefits particularly from the high-fidelity simulator environment, which preserves much of the complexity of actual operations but is also under instructor control. Performance assessment during revenue flights is also important, ideally, using performance measures from pilots and operational outcomes. In addition, judgments about the training that is provided by participating pilots can provide valuable feedback, particularly in the early stages of training program development.

Both formative evaluation, which guides training program development, and summative evaluation, which assesses the effectiveness of the program as implemented, are useful. Our emphasis has been more toward summative evaluation, as we expect these would be needed in evaluating whether any new training requirements are met. Nevertheless, current AQP training evaluation is conducted in the framework of continuous improvement, and we expect this would be true if new requirements, e.g., for monitoring, are introduced. Thus, ongoing, as well as initial evaluation, would likely be needed.

Where we have commented about how evaluation of monitoring training programs might fit into existing practice, our examples primarily related to AQP. AQP training is generally more flexible, more established, and more resourced. Thus, AQP will likely be the point of adoption of new training programs and their evaluation. We conjecture that new training practices, and responses to anticipated changes in FAA requirements, will first be developed and adopted in AQP programs and then spread, with appropriated adaptations.

Program evaluation should address training topics that likely will improve monitoring, namely, those aspects of performance that are both useful and possible to train. Some limitations cannot be trained away. For example, we do not believe that people are capable of sustained attention (vigilance) for long durations without lapses—occasional unintended shifts in attention will occur regardless of the training. Similarly, decrements to monitoring may stem from fatigue, stress, or illness, but the direct impact of these factors on monitoring cannot be 'trained away.'

We provided a general framework for structuring evaluation as well as identifying the types of content that should be evaluated. Several recommendations deserve emphasis:

- Conceptual understanding. This is vital to train and assess for current airline operations. Modern transport category aircraft and airline operations are complex and are dependent on generalizable knowledge, which can be widely applied. This provides, must be adapted, a basis for problem solving when procedures relevant to the situation do not exist or cannot be accessed. Reasoning and comparison are critical parts of monitoring.
- Exercises in simplified contexts. Assessment of skill and knowledge components in simplified contexts is a powerful part of program evaluation. This allows diagnosis of program strengths and weaknesses efficiently because it is feasible to use multiple, quick tests to assess broader application of knowledge than is feasible in a simulator.
- Assessment of integrated performance in high-fidelity simulators with realistic scenarios, and in revenue flights. Demonstration that training benefits performance in a complex, realistic context is a very important part of program assessment. Simulators can provide a combination of control and realism, this environment is recognized as vital for training, and it is also very valuable for program assessment. Improved pilot performance in revenue flight is a fundamental goal of a change in training and is correspondingly critical to assess.
- Safety indicators. Training is intended to improve (some aspects of) safety and relevant indicators of safety need to be identified. These indicators can be a measure of the training program effectiveness. The training objectives must be aligned with the safety objectives, and relevant safety indicators are required to measure performance against the safety objectives.
- Stress testing. In each type of assessment, evaluation activities should include tasks that "stress test" the monitoring skill(s) that were trained or use other methods to push pilot performance to its limits. Inclusion of surprising scenarios is an important tool for this objective.
- Assess actual effectiveness of training. The aim of evaluation is to measure whether the training causes improvement. Planning ahead for how data is collected can mean you get specific, accurate information about whether (and how much) the training program was the cause of any performance improvement. We also provided guidance regarding tradeoffs for evaluation design.
- Implementation challenges. Development of an effective training program evaluation for monitoring may benefit from changing training or staffing of instructors and evaluators. It will also benefit from strategic use and adaptation of existing data collection opportunities.

In addition, in Section 11 we provided a method overview of what combinations of evaluation context, materials, and tasks are likely to be highly informative but moderate in cost, thus providing good value.

12.2. Scope and Further Considerations

The framework for evaluating training programs provided here—for both evaluation content and structure—is necessarily general. We identified knowledge and skills underlying monitoring based on the Sensemaking Model of monitoring. We identified a large set of measures for the various monitoring competencies in varying assessment contexts across Kirkpatrick's Levels of Evaluation. We provide a toolkit and design strategies, not detailed recipes or blueprints. The general specification provides flexibility: it can be applied to assessing a range of training programs for monitoring, with different specific needs.

This research activity did not attempt to detail the specific content of training or of program evaluation. While there are certainly monitoring challenges that are encountered in most operations (e.g., managing trajectory and energy on descent), the details of training and training evaluation will need to fit with the needs of individual operators and fleets. For example, in addition to identifying what should be known at the end of training, the training gaps should be identified so that training can concentrate on the competencies that are not already mastered. Different pilot training groups may have different experience and correspondingly different training needs.

A specific evaluation of a program for training monitoring can use the Sensemaking Model of monitoring as the design framework. To fill in specifics of evaluation requires identifying where the greatest needs are, both with respect to: a) monitoring components (mental model, gathering evidence, task management, etc.) and b) topic (flight path, air space, systems, etc.), and context of monitoring (phase of flight, types of threats or pressures). Both training and evaluation of training will likely need to sample across most aspects but concentrate on those of most concern.
For example, using components and topics that are more familiar may be a good place to introduce concepts (perhaps monitoring state of aircraft systems in cruise) while more challenging situations would be evaluated in simulated flight (perhaps flight path/energy management in descent with multiple changes by ATC).

Relevant measures from this rich set we identified can be selected and tailored depending on the ability to capitalize on and to shape data collection used for other purposes within the organization. Coordinated data collection and data sharing can be a tremendous aid to managing cost. The same measures collected to evaluate pilot performance as part of training can contribute data to the evaluation of the training program. Input to pilot evaluation plans within the training program to aid program evaluation needs is very helpful. A further opportunity comes from coordination between evaluation of training programs and data collection for operational safety programs (e.g., FOQA, LOSA) conducted by the safety management group.

This report addresses evaluation of programs for training the skills and knowledge needed for *monitoring*. Very little prior work focuses specifically on training or evaluation of training for monitoring. Nevertheless, monitoring overlaps with various components called out in nontechnical skill training (e.g., CRM, TEM). Situation Awareness is a topic that is perhaps most central to our definition of monitoring, but elements such as Problem Solving and Task Management are relevant as well. A variety of sources provide guidance on training and evaluation of these more established elements. Existing FAA training guidance for AQP is provided in AC #120-54A (FAA, 2017) and ICAO (2013) provides particularly detailed guidance (especially Appendix 2). We do not address training for the overall role of pilot monitoring with its more inclusive set of functions. Comments about this role can also be found in ICAO (2013), section 1.7.3.

This report addresses evaluation of *training programs*. This is a different goal than assessing whether pilots have certain knowledge and skills. While these goals are related, the first goal concerns evaluating a process (of producing change) and the second concerns evaluating a state (however arrived at). If the goal is checking whether pilot performance is over some threshold, different methods are relevant. For example, if evaluating the state of individual pilots' performance, before and after comparison is not needed; if evaluating a process, testing every pilot on every competency is not needed.

Although we did not focus on training design, the detailed relation between content of training and content of program evaluation is important. Program evaluation should consider retention of taught material over time, generalization to new scenarios and problems not specifically taught, and transfer from training into operational environments. To do this, it is important in program evaluation to know whether a particular task or scenario was used in training. If it was, and is tested again in the same context, performance can assess retention, but not generalization or transfer. If a task and scenario was not used in training, this novel case, tapping into the same competencies, can evaluate generalization. If the task and scenario were used in training, but the task is now tested in a different context, it can evaluate transfer (e.g., from simulator to revenue flight). While retention, generalization, and transfer are matters of degree, they are very important elements of program assessment. This assessment requires careful consideration of the relation among pilot activities used in training, in assessment within training, and in program assessment.

Both pilots monitor, but it is a central duty of the PM. Training is often organized around roles. This report, however, addresses evaluation of training the monitoring function, not the role. If training were to be organized by role, training the PM role would clearly require coverage of additional PM functions, not just monitoring.

Our characterization of monitoring as sensemaking clearly places monitoring as part of the competencies called crew resource management in the United States and non-technical skills in Europe. The evaluative framework and suggestions we provide may be applicable to a broader scope than just monitoring, but we do not explore the range of applicability.

Finally, it is worth noting that we do not focus on training monitoring or awareness that results from noticing very salient events that "pull" attention. These "bottom-up" aspects of attention and awareness certainly impact monitoring and inform the situation model. In the context of the highly automated glass cockpit, however, we suspect these responses are quite difficult to change by training. Thus, we have not prioritized methods for evaluating the impact of a training program targeting these processes. Possible impact resulting from such a training program could be evaluated based on change in monitoring performance, using measures discussed here. Future developments will be of interest.

12.3. Future Work

This report provides a broad framework for evaluating training programs that are intended to improve pilot monitoring. The intended outcome of improved monitoring is safer operation. Several topics are logical next steps for research to improve training of monitoring for airlines, including these:

- Development of training programs for monitoring and of methods for evaluating such programs should develop hand in hand. The Sensemaking Model's analysis of monitoring skill can be used to analyze and build up this complex cognitive skill.

Careful consideration will need to be given to how monitoring relates to other skills, and implications for training.

  - Development of assessment of a novel training program implicitly or explicitly defines the content of training. That is, the training program will likely be designed to 'teach to the test' to ensure the program, and the pilots it trains, are approved.
  - The relation between the training and the training evaluation, however, is somewhat complicated. In particular, a program assessment can only measure a subset of the changes desired from the training program. On the one hand, it is desirable for the evaluation to be general enough that it can be applied to different implementations of training. Our framework here emphasizes generality. On the other hand, the relation between the specific content and methods used in training and used in program assessment are important. In the extreme, if the identical materials and tasks used in training were used in program assessment and the performance was found to be excellent, this might say little about training effectiveness for other situations or tasks. (Of course, program evaluation need not depend exclusively on direct measurement of pilot performance.) If training is procedure oriented and the assumption is that 'all' procedures can be trained for use in 'all' situations, generalization might not be a training goal. As complexity increases, it is increasingly clear that even in proceduralized domains such as piloting "procedure execution" is just one skill required for piloting. Effective monitoring, like decision making and problem solving, ideally is a rather general skill, and further is not tied to a small set of stereotyped situations.
  - Analysis is needed both of the intended scope of generalization of training and of the intended relation between the content of training and the content of program evaluation.
  - Parallel development means that the program evaluation has something to evaluate, and feedback about training successes and challenges can inform training evolution.

- Many useful measures of monitoring components and integrated monitoring performance exist at the level of impact on pilot performance and these can be tailored to the particular tasks and content needed. However, impact measures at the operational level deserve considerable development.
- Focused discussion with airlines should attempt to identify what situations or topics seem to pose the biggest monitoring challenges, to identify what core content is common across airline operations, and to explore the degree of commonality and difference across airlines and fleets.
- Development of training and of assessing that training, "in miniature," for a bounded set of situations that are known to be important is a good next step. For example, applying the Sensemaking Model of monitoring, in detail, to monitoring from top-of-descent briefing with a particular focus on monitoring and managing complex ATC clearances might be a focus for development. A matched program assessment for this limited domain could be developed in parallel.
- Developing a library of scenarios that pose particular monitory challenges is an important resource for research and for operational use. These can be specified abstractly to allow implementation on a variety of fleets, and also have accompanying cockpit video of pilots flying the scenarios and exhibiting some variety of monitoring behaviors.

- Empirical, lab-based testing of a small, newly developed training program would be valuable; it should include comparison of efficiency and effectiveness of different training exercises. In turn, this could be used in development of program assessment.
- Investigating how both the monitoring training program and the evaluation fit within existing practices and regulations.
- Training design may focus on monitoring as a skill needed in every role, or might have distinct, role-based training, e.g., for the PM role.

References

Billman, D., Mumaw, R. J., & Feary, M. (2019). *Best Practices for Evaluating Flight Deck Interfaces for Transport Category Aircraft with Particular Relevance to Issues of Attention, Awareness, and Understanding CAST SE-210 Output 2.* NASA/TM-2019-220390.

CAST. (2014). Airplane state awareness joint safety analysis team. https://www.skybrary.aero/bookshelf/books/2999.pdf.

Durso, F. T., & Dattel, A. R. (2004). SPAM: The real-time assessment of SA. In S. Banbury & S. Tremblay (Eds.), *A Cognitive Approach to Situation Awareness: Theory, Measures and Application.*

Endsley, M. R. (1995). Measurement of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 65–84. https://doi.org/10.1518/001872095779049499

FAA. (2017). *Advanced Qualification Program AC No: 120-54A* (AFS-200 AC No: 120-54A).

Flight Safety Foundation. (2014). A practical guide for improving flight path monitoring. Alexandria, VA: FSF.

Flin, R., Yule, S., Paterson-Brown, S., Maran, N., Rowley, D., & Youngson, G. (2007). Teaching surgeons about non-technical skills. *The Surgeon, 5*(2), 86–89. https://doi.org/10.1016/S1479-666X(07)80059-X

Flin, R., Martin, L., Goeters, K.-M., Hormann, H.-J., Amalberti, R., Valot, C., & Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. *Human Factors and Aerospace Safety, 3*(2), 95–117.

Flin, R., O'Connor, P., & Mearns, K. (2002). Crew resource management: Improving team work in high reliability industries. *Team Performance Management: An International Journal, 8*(3/4), 68–78. https://doi.org/10.1108/13527590210433366

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models* (7. pr.). Erlbaum.

Hoermann, H.-J., Banbury, S., Blokzijl, C., Dudfield, H., Lamers, J., Lehmann, O., Lodge, M., & Soll, H. (2003). *Experimental Validation ESSAI/DLR&Q_Q/WPR/WP5/2.0—Status:* E S S A I -Enhanced Safety through Situation Awareness Integration in training.

Hoffman, R. R., Ward, P., Feltovich, P. J., DiBello, Lia, Fiore, S. M., & Andrews, D. H. (2013). *Accelerated expertise: Training for high proficiency in a complex world.* Psychology Press.

Holt, R. W., Boehm-Davis, D. A., & Hansberger, J. T. (2001). *Evaluation of Proceduralized CRM at a Regional and Major Carrier.* https://www.researchgate.net/publication/237580032_Evaluation_of_Proceduralized_CRM_at_a_Regional_and_Major_Carrier

ICAO (International Civil Aviation Organization). (2013). *Manuel of Evidence-based Training*.

Jones, D. G. (2000). Subjective Measures of Situation Awareness. In Endsley, Mica R. & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement*. Lawrence Erlbaum Associates.

Kieras, D., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science, 8*, 255–273.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed). Berrett-Koehler.

Koteskey, R. W., Hagan, C., & Lish, E. T. (2019). Line Oriented Flight Training. In *Crew Resource Management* (pp. 283–322). Elsevier. https://doi.org/10.1016/B978-0-12-812995-1.00010-5

Landman, A., van Oorschot, P., van Paassen, M. M. (René), Groen, E. L., Bronkhorst, A. W., & Mulder, M. (2018). Training Pilots for Unexpected Events: A Simulator Study on the Advantage of Unpredictable and Variable Scenarios. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 60*(6), 793–805. https://doi.org/10.1177/0018720818779928

Mumaw, R.J., Billman, D., & Feary, M. (2019a). Factors that Influenced Airplane State Awareness Accidents and Incidents. CAST SE-210 Output 2. NASA/TM-20205010985.

Mumaw, R. J., Billman, D., & Feary, M. (2019b). Identification of Scenarios for System Interface Design Evaluation. CAST SE-210 Output 2 Report 5 of 6. NASA/TM-20205010988.

Mumaw, R. J., Billman, D., & Feary, M. (2020). Analysis of Pilot Monitoring Skills and a Review of Training Effectiveness. NASA/TM-20210000047.

Mumaw, R. J., Sarter, N. B., Wickens, C. D., Kimball,S., Nikolic, M., Marsh, R., Xu, W., & Xu, X. (2000). *Analysis of Pilots' Monitoring and Performance on Highly Automated Flight Decks* (Final Project Report NASA Ames Contract NAS2).

O'Connor, P., Hörmann, H.-J., Flin, R., Lodge, M., Goeters, K.-M., & JARTEL Group, T. (2002). Developing a Method for Evaluating Crew Resource Management Skills: A European Perspective. *The International Journal of Aviation Psychology, 12*(3), 263–285. https://doi.org/10.1207/S15327108IJAP1203_5

Reis, H. T., & Judd, C. M. (Eds.). (2009). *Handbook of research methods in social and personality psychology* (Reprint). Cambridge Univ. Press.

Sarter, N. B., & Woods, D. D. (1995). How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 5–19. https://doi.org/10.1518/001872095779049516

Stewart, M., Matthews, B., Janakiraman, V., & Avrekh, I. (2018). Variables Influencing RNAV STAR Adherence. *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC),* 1–10. https://doi.org/10.1109/DASC.2018.8569220

Vallacher, R. R., & Wegner, D. M. (1987). What do people think they're doing? Action identification and human behavior. *Psychological Review, 94*(1), 3–15. https://doi.org/10.1037/0033-295X.94.1.3

Visser, S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey Research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology.* Cambridge University Press.

Appendix A. Alternative Modeling Perspectives

Many models of cognitive processes and of cognitive work emphasize a linear sequence of steps, running from perception to action. This type of model may be more familiar, and it can be applied to monitoring, as shown in Figure A-1. Figure 8a comes from an earlier characterization we provided of monitoring. This shows a sequence of steps, and also indicates the components that are part of monitoring. This model recognizes the importance of feedback but interaction among processes (or with the current understanding) are not emphasized. The Sensemaking Model includes but reorganizes these components, emphasizing the pervasive role of comparison and situation understanding throughout the processes of monitoring. Figure A-2 provides a rough illustration how linear processes may be included in the cyclic model. This is a rough illustration and the mapping is not precise.

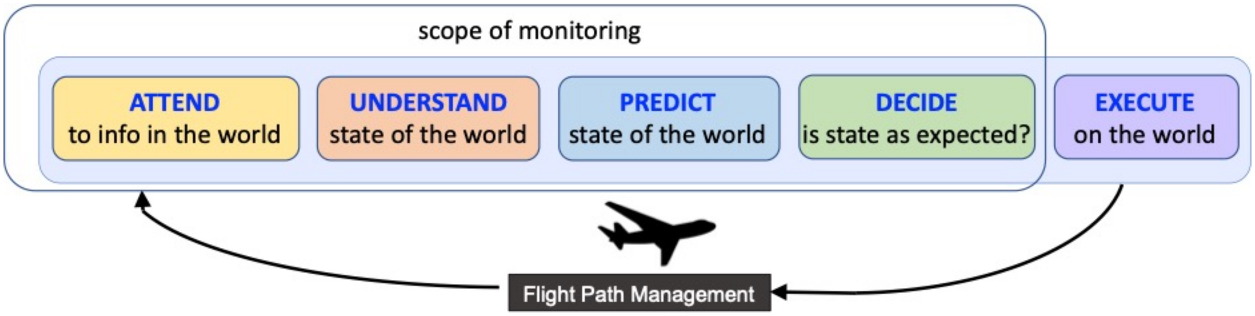

Figure A-1. Linear models emphasize a sequence of steps from attending through thinking and action.

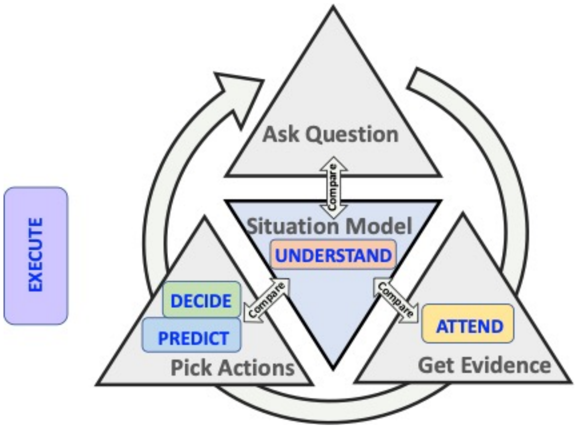

Figure A-2. The processes in the linear model are included in the Sensemaking Model but are organized differently.