NASA/TM—2019–220390

# 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
Report 6 of 6

Dorrit Billman
*San Jose State University Foundation*

Randall J. Mumaw
*San Jose State University Foundation*

Michael S. Feary
*NASA Ames Research Center*

March 2019

# 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 organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

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 to help@sti.nasa.gov

- Phone the NASA STI Information Desk at 757-864-9658

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

NASA/TM—2019–220390

# 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

Report 6 of 6

Dorrit Billman
*San Jose State University Foundation*

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*

March 2019

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

# Acronyms and Definitions

AC.....Advisory Circular
ASA .....Airplane State Awareness
ATC .....air traffic control
CAST .....Commercial Aviation Safety Team
CFR.....Code of Federal Regulations
ECG .....electrocardiography
EEG.....electroencephalography
ERP .....event-related potential
FAA .....Federal Aviation Administration
fMRI .....functional magnetic resonance imaging
fNIR or fNIRS .....Functional Near-Infrared Spectroscopy
ft.....feet
GSR.....galvanic skin response
HCIPA .....Human-computer Interaction Process Analysis
HF .....human factors
IDA .....intentional discriminative action
LOC .....loss of control
MATLAB .....Matrix Laboratory
MEL.....minimum equipment list
NAS .....National Airspace System
NASA .....National Aeronautics and Space Administration
PARC.....Performance-based operations Aviation Rulemaking Committee
PM.....pilot monitoring
RAFIV .....Reformulate, Acess, Format, Insert, Verify
SA-SWORD .....Situation Awareness – Subjective Workload
SAGAT.....Situation Awareness Global Assessment Technique
SART .....Situational Awareness Rating Technique
SE.....Safety Enhancements
SPAM .....Situation Present Assessment Method
TOD .....top-of-descent
VNAV PTH .....vertical navigation path
VNAV SPD .....vertical navigation speed

# 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
## Report 6 of 6

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

## SE-210 Project Overview

The Commercial Aviation Safety Team (CAST) created a team to analyze a set of incidents and accidents associated with the flight crew's loss of awareness of aircraft attitude or energy state. These events are referred to more broadly as a loss of Airplane State Awareness (ASA), and they are a substantial subset of loss of control (LOC) accidents. A subsequent CAST ASA team developed a set of mitigation strategies—referred to as Safety Enhancements (SEs)—to reduce the likelihood of ASA events occurring in the future. Six of the SEs (SE 200, 207 through 211) requested further research on mitigation strategies. Our work was specifically intended to address research identified in SE 210 Output 2 (see https://www.skybrary.aero/bookshelf/books/ 2540.pdf).

SE-210 Output 2 addresses the contributions from the flight deck interface in shaping pilot awareness. More specifically, the focus is on assessing or *evaluating the flight deck interface to determine how well it supports ASA*. We have produced a series of reports on this topic.

1. In a report titled "Overview of Research Approach and Findings" we introduce our research approach and compile our key observations and findings. This provides a summary of how our research method developed and what we found.
2. Part of our work was a more-detailed analysis of the role of awareness in the ASA events. In a report titled "Factors that Influenced Airplane State Awareness Accidents and Incidents" we describe a number of factors that contributed to the apparent loss of awareness, or to the resulting loss of control. This analysis demonstrates that pilot attention and understanding of the system are important elements of awareness. This report also offers proposals for modifications of the interface to mitigate those factors, and then, describes how you might evaluate the effectiveness of those proposed modifications.
3. In a related report, titled "The Role of Alerting System Failures in Loss of Control Accidents," we analyze how alerting for LOC-related hazards, such as low airspeed, unreliable airspeed, and approach to stall, can fail to lead to an upset recovery. Alerting is the last line of defense against flight path management hazards; it is there to ensure awareness when pilot-driven attention and awareness fail. This report looks at why alerting does not always save the day.

Through our work, we had the opportunity to become more familiar with current evaluation and certification rules, guidance, and practices that define the process for the applicants (equipment manufacturers) and the Federal Aviation Administration (FAA). Evaluation and certification of

flight deck interface elements consider a broad range of flight crew performance topics. We narrowed the focus of our work to flight crew awareness, attention, and understanding, and specifically examined these aspects of human performance in relation to relevant rules (e.g., 14 CFR 25.1302) and advisory material (e.g., AC 25.1302-1). This new material offers a more complete description of flight crew performance issues in the context of the flight deck interface; however, no consistent approach for application has been established.

4. In a report titled "Evaluation Issues for a Flight Deck Interface," we attempt to describe the broader scope of flight crew performance issues to show how awareness and attention issues fit within the larger set. We also do an inventory of FAA certification rules to demonstrate that there are not rules that apply to every issue. AC 25.1302 has improved guidance for addressing evaluation of awareness, attention, and understanding, and we hope that our work can contribute to future updates of the guidance material.
5. A related report, titled "Identification of Scenarios for System Interface Design Evaluation," focuses on the operational scenarios that can be used in the context of interface evaluation. It offers several perspectives on how to ensure that pilot or flight crew performance is evaluated in an important operational context. Because it is unlikely that evaluation can be performed for the full range of operational settings, this report offers a method for selecting appropriate scenarios.

Finally, the bulk of our work in this project was focused on methods for evaluating a flight deck interface for how well it supports awareness and its critical elements: attention and understanding.

6. The current report, titled "Best Practices for Evaluating Flight Deck Interfaces for Transport Category Aircraft with Particular Relevance to Issues of Attention, Awareness, and Understanding," focuses on evaluation techniques and metrics. It considers opportunities to evaluate the interface from early to late stages of development; it considers the various ways in which the interface can fail to support awareness, attention, and understanding; and, it summarizes appropriate evaluation methods for different issues. This report draws on the characterization of issues and of scenario selection presented in other reports that are relevant to awareness.

## 1. Motivation and Claims Underlying this Report

Recent safety studies have shown that loss of awareness or mismanagement of flight crew attention can lead to upsets and accidents (Mumaw, Haworth, & Feary, 2018; FAA: Performance-based operations Aviation Rulemaking Committee, 2013). The flight crew's processes of attention, awareness, and understanding are critical to safe flight. The flight deck interface plays a key role in adequately supporting flight crew attention, awareness, and understanding. The sound evaluation of the support provided by the flight deck interface, therefore, plays an important role in promoting safety. However, there are open questions about what evaluation methods are most effective and efficient for the task of assessing how well an interface or interface component supports attention, awareness, and understanding. An understanding of evaluation methodology can aid identifying and adapting those methods that will make the best use of available resources to provide the most relevant information. This report describes available, sound evaluation methods for assessing the flight deck with respect to providing support for attention, awareness, and understanding.

Several key claims are the foundation of the report.

- *Smart use of limited resources*. Understanding what methods are most effective and efficient for addressing a particular set of issues is very valuable. Understanding this helps design an evaluation that provides highly relevant information while making the best use of available resources.
- *Issue oriented*. Systematic identification of issues or concerns with respect to attention, awareness, and understanding enables evaluation resources to be focused on the aspects that seem to present the greatest threat to safety. Issue identification helps to identify relevant methods as well as relevant compliance rules.
- *Early evaluation start*. Starting evaluation in the early phases of the design and development process of a flight deck product contributes to an effective, efficient evaluation process. Early-phase evaluation means that issues can be identified early, when the cost of change is low. In addition, early-phase methods themselves are often lower cost. The opportunity for low-cost modifications to resolve issues is a key reason why early involvement of a regulator is advantageous to an applicant. Importantly, early evaluation fits within an overall process of assessment across multiple phases of development.
- *Performance-based measures*. Methods that capture operator performance are valuable for evaluating many issues. They are, however, particularly valuable for issues concerning attention, awareness, and understanding. It is very difficult to assess attention, awareness, and understanding without data from an operator measured while carrying out meaningful tasks with the equipment being evaluated. A variety of methods for doing this are available, including low-cost methods targeting a very specific issue.
- *Adapting methods to the case at hand.* Both the available resources and the likely risks vary from one evaluation to another. This means methods will need not only to be selected but also modified to suit the case. This report aims to describe: a) general methods useful for evaluating support for attention, awareness, and understanding; b) guidance about conditions when the methods may be particularly useful; and c) illustrations of how methods might be applied in different evaluation phases to address different issues.

## 2. Purpose of this Report

This report offers potential guidance and methods of compliance for determining the acceptability of the flight deck interface and its components in regard to attention, awareness, and understanding. Two themes organize the guidance we offer. First, the evaluation is more likely to be effective and efficient when evaluation begins early in design and development, and when regulators are engaged early in the process. Second, we provide descriptions of methods useful through the phases of development. Application of these evaluation methods is guided by the issue or issues that are the evaluation concern. Many issues will benefit from evaluation at multiple points in the development process and we describe evaluation methods available in each development phase. We hope this will help design evaluation plans appropriate to the particular flight deck or flight deck components and to their associated evaluation issues. For brevity, we refer to a flight deck interface or interface component undergoing evaluation as a device. The companion report titled “Evaluation Issues for a Flight Deck Interface”(Mumaw, Haworth, Billman, & Feary, 2019) provides broad guidance on recognizing evaluation issues and the mapping between an issue and the regulatory basis.

We focus on evaluating how well a device supports attention, awareness, and related understanding, though we hope aspects of this report will be applicable to a wider range of human factors issues.

We focus on these topics because issues here are frequent contributing causes to recent aircraft energy state awareness incidents and accidents (Mumaw, Billman, & Feary, 2018; Mumaw, Haworth, et al., 2018; Performance-based Operations Aviation Rulemaking Committee [PARC], 2013; The Airplane State Awareness Joint Safety Implementation Team [CAST, 2014]). Further, the importance of supporting these and other higher level, cognitive functions will increase for future automated systems in the airspace. The complexity and integration of future systems will increase, and this in turn will impose more challenging cognitive work on pilots, as well as on the other operators in this space.

Aircraft certification and approval processes are both highly variable and complex, particularly for issues relevant to supporting improved attention, awareness, and understanding. As a result, direct, routine application of evaluation methods is unlikely. Rather, methods will need to be tailored to specific evaluation projects. Thus, we present method fundamentals, their strengths and weakness, and also provide illustrations of issues that may be addressed with methods available at different phases of development. We aim to provide guidance about the most accurate and efficient methods of assessments. This provides a basis for assessing relevance of data that an applicant provides and hence its adequacy for determining acceptability of the device.

We do not address what issues should be assessed, as this is a function of the device undergoing evaluation. We do not address the extent of assessment needed for any issue, as this is a function of the complexity, novelty, integration, and all aspects concerning safety of a device. For devices with a very low safety risk, less intensive investigation will be warranted. We do not address how much data is sufficient to resolve an issue for different cases.

# 3. Attention, Awareness, and Understanding

## 3.1. Overview

Most complex actions depend on awareness. Awareness refers to conscious, reportable knowledge. What is in awareness typically depends on what is attended. Attention is limited, so people must dynamically reallocate attention to have adequate awareness in any complicated situation.

A key human strength to compensate for limits on attention and awareness is our ability to build up an integrated understanding of the situation. Mental models and situation models are an important part of this understanding. *Mental model* refers to a person's general understanding of how something works or a type of situation. *Situation model* refers to the person's representation of the specific, dynamic situation. A situation model is typically based on a mental model of the situation type. Situation models are more specific, include details about the current case, and are updated dynamically as the current situation changes. Once attended information is understood and linked to a dynamic situation model, the information can be recalled more reliably and better used in assessment and decision making.

Design of flight deck components should respect and support attention limitations and should aid understanding. Evaluation of a flight deck device should assess how well it supports attention, awareness, and understanding.

Critically*, performance-based data* are needed to evaluate how well the flight deck supports pilot awareness. People do not have good intuitions about what, and specifically how much, information will enter awareness, but systematically overestimate what will be noticed (Levin, Momen,

Drivdahl, & Simons, 2000). Rather than relying on intuition, even that of experts, sound assessment critically requires performance data that is tailored to the issue in question.

Guidance on collecting relevant performance-based, as well as useful inspection-based data, is a major part of this report. The methods in this report draw on research methods for investigating attention, awareness, and understanding; the methods are adapted to evaluating how adequately a flight deck device will support pilot attention, awareness, and understanding.

### 3.2. Limited Attention and Its Allocation: Supporting Stimulus-driven and Expectation-driven Allocation

The flight deck interface has two functions in helping the pilot allocate attention. It needs to "push" important information to a pilot whether or not the pilot is looking for that information; here attention is stimulus-driven. The flight deck also needs to help a pilot easily "pull" the information they are looking for; here attention is driven by pilot expectation. The evaluation plan should assess the interaction design for supporting both external, stimulus-driven (exogenous) and internal, expectation-driven (endogenous) aspects of attention. Of course, both aspects of attention are usually in play, in varying degrees.

Flight deck interfaces should provide external stimuli designed to shift the pilot's attention and "push" information to the pilot when needed. Alerting is the clearest case of pushing information to a pilot, in this case about the existence and nature of an important change, but "pushes" may come from other sources and concern other information as well. Alerts are particularly valuable when the alerting system can identify that the alerted information is indeed the highest priority for the pilot to attend to. Because attention is limited, pulling attention to one information source means attention to previously attended information is reduced and other tasks disrupted. Thus, there are limits to the extent that awareness can usefully be directed by highly salient alert stimuli, and alerts designed to grab attention must be used very selectively. Although alerting provides the clearest examples of "pushing information to the pilot," the need for directing attention is widespread, as in highlighting values that have changed recently.

Critically, the interface should also support the pilot's current goals, helping the pilot "pull in" the information they are looking for to support the current task. Awareness of the values of important, relevant variables allows the pilot to maintain an appropriate, accurate model of the situation, to make appropriate decisions and take the relevant command actions. In turn, this gives pilots the best possibility of keeping the aircraft performing safely. Monitoring is primarily expectation-driven, as pilots pull in the information they are seeking, to update their integrated situation model (sometimes called a mental model).

Endogenous factors concerning internal goals and expectations are remarkably powerful in guiding attention and hence awareness. People are good at noticing changes or information they are looking for or expect and poor at detecting information they are not looking for. Indeed, people screen out information not relevant to their current goal to concentrate on the goal activity. People can look directly at something and not be aware of it if they are not looking for it; similarly, they can be directly looking at an area but be unaware of changes there. This *inattentional blindness* phenomenon is widespread (Neisser & Becklen, 1975; Simons & Chabris, 1999), even among experts (naval technicians in DiVita, Obermayer, Nugent, & Linville, 2004; radiologists in Drew, Võ, & Wolfe, 2013; pilots in Fischer, Haines, & Price, 1980; multiple medical disciplines in Lum, Fairbanks, Pennington, & Zwemer, 2005) and implications for applied settings have been reviewed

(Durlach, 2004; Varakin, Levin, & Fidler, 2004). Attention limitations of this type are a key factor driving performance in many tasks. Further, people have very poor intuitions about these phenomena (Levin et al., 2000), making objective performance assessment particularly critical.

### 3.3. Attention, Awareness, and Understanding Issues from the Flight Deck

We group issues for consideration within four broad topics concerning attention, awareness and understanding. The four topics address how well the device carries out these interrelated functions:

- provides information the pilot is looking for
- ensures accessibility of information provided
- directs pilot attention to information about important changes
- supports situation understanding and assessment of action

Broadly, these four topics are ordered from "lower level" to "higher level" in that many of the "higher level" functions depend on the "lower level" functions. These topics can be evaluated at the several evaluation phases, described in the next section. Example applications organized first by phase of evaluation and then grouped by these four topics are presented in Appendix A. Further discussion of human factors issues relevant to evaluation are provided in Mumaw et al. (2019).

## 4. Framework for Evaluation

### 4.1. Flow of Evaluative Activities from Early in Design Specification

How evaluation is conducted across the phases will depend very much on the nature of the device being evaluated, such as its complexity, novelty, and integration. Different issues will merit more intense evaluation depending on the device.

We divide the evaluation process into four phases of evaluation. The boundaries between phases are not sharp, and phase of development for different aspects of a device may differ as well. The purpose of a phase-based approach is to help design an evaluation plan that will identify and remedy problems as early as possible, thus reducing cost of both the evaluation and the overall certification process. The four evaluation phases are:

1. Design review
2. Prototype performance
3. Simulator-based testing
4. Flight testing

Table 1 shows evaluation phases. The evaluation phases differ in the materials available for evaluation and the tools needed for the methods applicable in each phase. Earlier evaluation phases require less fully-developed materials and equipment. Methods useful in one phase can also be used in any later phase.

Table 1. Characteristics of Methods in the Four Evaluation Phases

| *Evaluation Phase/Method* | | | *Materials Available for Evaluation* early ----------> late | | | |
|---|---|---|---|---|---|---|
| | *Purpose* | *User Tasks* | *Design Documents* | *Prototype** | *Simulators* Mixed Levels* | *Built Device* |
| Inspection-based | | | | | | |
| **Design review** | Formative | ✓ | ✓ | | | |
| Performance-based | | | | | | |
| **Prototype** | Formative | ✓ | | ✓ | | |
| **Simulation** | Formative & Summative | ✓ | | | ✓ | |
| **Flight Test** | Summative | ✓ | | | | ✓ |

**developed sufficiently for issue tested*

Two high-level factors characterize the evaluation phases: *Inspection-based* versus *Performance-based* methods and *Formative* versus *Summative* evaluation goals. Concerning Inspection-based versus Performance-based, Inspection-based Methods depend on an expert inspecting a device or representation of the device, without actually carrying out the tasks the device is supposed to support; thus it need not be possible to actually 'do the work' for an informative inspection, and it is possible to apply when only design documents are available. Performance-based Methods require an operator to carry out some goal-defined task and for data to be collected about how the tasks were performed. Thus, at least some form of prototype is needed for the operator to work with. The tasks on which performance is measured, however, can be small parts of the eventual scope of work supported; thus performance-based methods can be used quite early. Further, performance-based measures using targeted, low-fidelity prototypes can be quite inexpensive.

If the device being evaluated is similar to existing devices in relevant aspects, historical performance data may be useful as well. Specifically, if certain aspects of a device are similar to an existing device with relevant performance data, additional performance data may be focused on the novel aspects of the device rather than aspects for which there is existing, relevant performance data.

Concerning Formative versus Summative goals, formative evaluation is concerned with identifying and diagnosing problems to be addressed and heavy use of formative methods is appropriate from early in design until design freeze. Summative evaluation addresses how well a product, which is intended to be final, in fact performs. While the end result of the evaluation process is summative, formative evaluation is typically critical for ensuring a successful outcome in a later, summative evaluation. This is particularly true for devices which are more novel, more complex, and more interactive with the context of use.

This report describes informative, efficient methods available at different points in the development process. Which of the methods to use and the stage of development to use them will be influenced by tradeoffs among many factors. For example, if particular aspects of the device are very similar to existing, successful devices, an early inspection method and little performance testing with

prototypes or simulator, followed by final confirmation that the device functions as designed may be appropriate. Conversely, evaluation of highly novel, complex, or integrated systems may greatly benefit from considerable performance-based testing. The suggestions here provide a basis for assessing tradeoffs between risks and evaluation costs. The methods and suggestions here are extensive but not complete: there may be ways of doing a useful, early assessment not identified here, or reasons for deferring assessment of a specific known issue until later.

## 4.2. Preview of Four Evaluation Phases

### *4.2.1. Design Review*

Systematic design review is likely to be most helpful at the point when design narratives and diagrams show or describe key elements of the interface appearance and behavior. A key purpose of an early design review is identification of issues that warrant further evaluation, perhaps with a more detailed design review method targeting the particular issue, or later-phase, performance-based evaluation.

The early materials that provide the input to a design review may be abstract or incomplete but still provide important information about the design. The design can be inspected to look for issues likely to result in inadequate support for human performance. For example, the hierarchical menu structure for flight plan information might be written down (inspectable), but not prototyped in a way that participants could try navigating from one type of information to another. Once someone can "do a task," performance-based, prototype methods maybe most useful. The design review and the Prototype Performance phases may overlap, and different aspects of the device might be developed on different schedules.

An inspector can take a careful look at the design documents and assess how information is organized, the nature of controls, and how this will support the flow of activity, both for carrying out the intended function of the device and for impact on other activities. Ideally, someone not involved in creating the design but with domain and HF (human factors) expertise should be the primary inspector.

The purpose of early review is to provide early and hence low-cost notice of HF problems. The most concrete "formal" findings from a design review are usually a description of the problems identified. Socializing the value of early detection of problems may be beneficial, perhaps from cases where this in fact saved the Applicant money. Applicants with HF staffing may well conduct this type of design review internally but bringing in FAA HF expertise and sharing results with the FAA will add value. Informal outcomes include alerting FAA to the developing project and resources its evaluation may need, as well as more informal concerns or opinions conveyed from the FAA to the Applicant.

While design review of a complex system is still valuable, performance-based methods become increasingly important as the device or the issues become more complex. Further, performance-based evaluation is particularly valuable, because it can be very hard to estimate from inspection what an operator's awareness will be and how it may vary depending on the situation.

### *4.2.2. Prototype-based Evaluation*

The Prototype-based phase of evaluation requires that the design has been instantiated in a static or dynamic interface, sufficient to be used in some task. In this phase it is possible to begin collecting (part-task) performance data. Performance data comes from a goal-driven task where someone is

measured carrying out the task. The task specified may be quite simple, such as identifying and reporting the value in some indicator, or quite complex. The type of users needed will vary depending on the issue being assessed, but, unlike inspectors, the users need not be human factors or design experts. Indeed, simple tasks such as finding or reporting values may not need any special qualifications of users. The amount of prior aviation training and training on the new device that are needed will differ depending on the issue. Specifically, useful data about very simple perceptual judgments in early prototyping may be gained from participants without prior flight experience. Higher-level issues such as projecting future state of the aircraft will require knowledgeable participants, specifically, pilots. Just as this phase may overlap with design review, evaluation using prototypes and using simulators may overlap: it may be possible to carry out virtually the same test on a well-developed prototype and a low-fidelity simulator.

Bench testing fits within this phase of evaluation. Bench testing assesses a physical component in isolation. It provides a good "prototype" of this component, but does not provide additional context, as part-task prototypes may do.

At the point of testing with prototypes and low-fidelity simulations, many issues concerning how well the device supports its intended functions, in isolation, can be assessed. Depending on the prototype, assessing performance in the context of the flight deck and across a wider range of task may be possible as well. Design of scenarios, that is the situations and activities used in evaluation, are important in performance-based evaluation, from prototype-based evaluation onward. If problems are identified from performance on tasks using a prototype, they can be addressed when the cost of a design change is quite low. Unaddressed problems are very likely to worsen when the device is used in a more complex context. Thus, as with design review, early evaluation can identify problems but early evidence that the device provides an acceptable level of support will need some confirmation for the more fully developed system. Later tests can be designed so they include coverage of activities that previously had problems.

Performance-based evaluation as soon as prototypes become available may be particularly valuable for identifying attention, awareness, and understanding issues. While some major problems (such as the failure to provide needed information) may be identified by inspection, it is very hard to predict impact on awareness. Designing efficient and effective prototype-based or simulation-based studies requires care to ensure that the performance data collected do address the intended issue for evaluation, thus ensuring the most efficient evaluation process.

### *4.2.3. Simulator-based Evaluation*

In this phase of evaluation, development has proceeded to the level where the component being evaluated can be used in the context of many or all relevant components of the flight deck and the simulator can dynamically provide changing states to the pilot to simulate flight. For at least part of the activities in this phase, the participants need to be pilots who are representative of pilots who will be using the equipment. They should not be test pilots or pilots involved in designing the device.

A critical part of simulation-based evaluation is design of scenarios. In addition to sampling scenarios using typical tasks in normal and non-normal conditions, scenarios should be designed to focus on "edge cases." Edge cases are situations with unusual combinations of conditions, particularly situations where interactions among events or system components might be revealed. For example, a normally useful audio alarm might be masked by another, the need to configure displays to show one type of information might conflict with an unusual information requirement,

and problems might appear with generic stressors such as high workload. These cases may be particularly likely to reveal design weaknesses. Further guidance on designing scenarios useful for performance-based evaluation, and particularly for prototypes and simulators, can be found in Mumaw, Billman, & Feary (2019).

Simulation-based evaluation provides the most powerful testing environment. Depending on the simulator level, this testing provides capabilities close to those in flight-testing. In addition, it allows assessment of dangerous scenarios as well as a range of situations which would be infeasible (e.g., weather conditions) or expensive (e.g., repeated landings) to test in flight. The value of simulator testing is very high, and data useful for many issues may be possible to collect from a single well-designed assessment on a low fidelity simulator.

#### *4.2.4. Flight Test*

Flight testing provides a final, summative evaluation. If prior phases of evaluation have been successfully carried out, no problems in design should be discovered in flight testing. Indeed, flight test is a very difficult setting to identify problems in attention, awareness, and understanding. Rather, flight testing can assess, over a limited range of conditions, whether the aircraft functions as designed. Flight testing is expensive, potentially dangerous, and limited in the range of safety-critical issues that are feasible to assess. Further, any safety problems identified at this point are extremely expensive to correct and the feasible solutions are much more restricted than if the problem were identified early in development.

### 4.3. Why This is Important and What We Cover

Currently there is not much guidance for evaluating many of the issues related to attention, awareness, and understanding. Advisory Circulars may provide very broad or generic advice for these topics. Thus, laying out relevant evaluation issues and methods may be helpful for the topics we consider. More detailed mapping between issues and existing rules is provided in Mumaw, Haworth, Billman, & Feary (2019). Building an evaluation plan that spans the phases of development has high value for issues concerned with awareness. Some problems can be identified early in design review (Phase 1), but prototype evaluation (Phase 2) is particularly valuable for performance-based, low-cost identification of problems with particular displays and alerts, prior to assessment in the more complex, dynamic environment of a simulator (Phase 3). We outline how a phase-based approach might be applied to the evaluation of issues concerning attention, awareness, and understanding.

### 4.4. Determining Scope of Evaluation

This report does not address how to set the appropriate scope or intensity of the evaluation process for certification or operational approval. Rather, we offer guidance about the most efficient methods for discovering and addressing a variety of issues throughout the phases of development. Identifying key issues as early as possible and prioritizing these issues in evaluation helps design the most efficient and effective evaluation.

# 5. Device Evaluation for Acceptance Testing: Method Fundamentals for Attention, Awareness, and Understanding

## 5.1. Overview

An understanding of methodology fundamentals is valuable because this lets an evaluator select and adapt the methods that efficiently address the issue of concern. We introduce some of these fundamentals and give examples of method choices that are feasible at different phases of development. Our examples focus on issues concerning attention, awareness, and understanding.

The methodology for any evaluation can be specified by the type of data, tasks, situation, people, criteria, and overall strategy. Table 2 provides an overview.

Table 2. Six Aspects of Evaluation Methods for Acceptance Testing

| | |
|---|---|
| Type of data collected | Data types, and their specific measures, differ in the type of information collected and have different strengths and weaknesses in what they can contribute to the evaluation of a device undergoing certification. |
| Tasks | The evaluation scenario includes the task or activity to be carried out and "what" the tasks are applied to, e.g. design sketches or a simulation of an aircraft. |
| Situations | The evaluation scenario also includes the situations that set the content and context of the tasks to be carried out; they may be very simple or highly naturalistic. |
| People | The people that generate data are inspectors (assess materials) or participants (perform tasks). |
| Criteria | In the context of certification, summative evaluation criteria specify how the device is judged to be "safe enough," whether by relative comparison to existing systems or to an agreed upon standard. |
| Overall evaluation strategy | Together, the combination of the types of data, tasks, situations, and people needs to be specified so that the collected data do in fact address the issue(s) in question. Further, considering the issues of concern together enables building an efficient evaluation plan: collecting the data that provides information about the greatest safety risks for the least cost. |

## 5.2. Data: What Types of Data and Measures are Collected?

### *5.2.1. Overview*

Several types of data (and their associated measures) may be used in certification-related assessment of aircraft flight-deck devices. Using a data type (such as verbal report "in the moment") for a particular evaluation requires identifying the specific measures of that data type to be used in that evaluation (such as what is to be reported, how responses will be scored). The specific measures used in a particular assessment are also called "dependent variables." As shown in Figure 1, the data types relevant to evaluation fall into three very broad categories:

1. Observations from inspection
2. Performance measures from carrying out tasks
3. Data from modeling

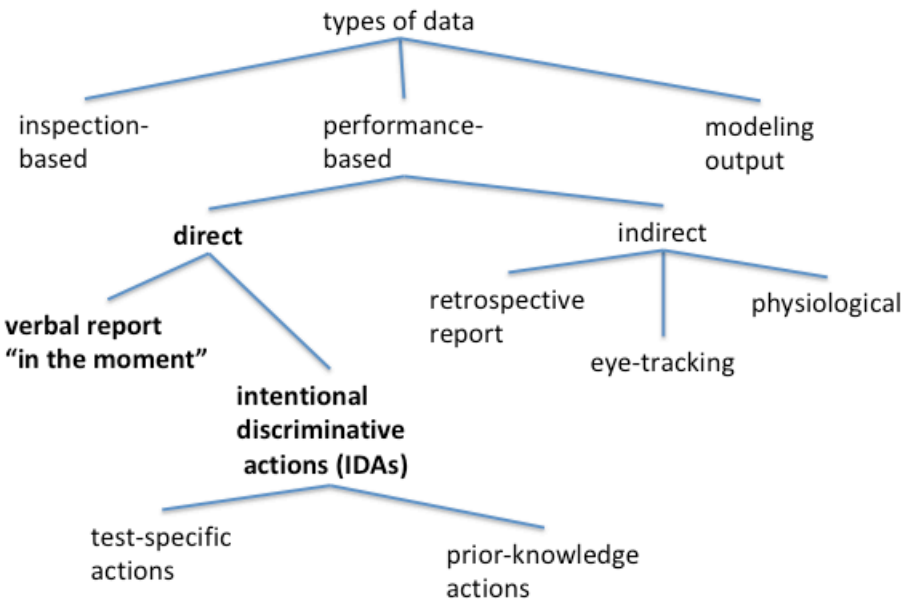


*Figure 1. Types of data useful in evaluations concerning support for attention, awareness, and understanding. Performance-based data, particularly **direct** performance data, is typically most informative. Also, designing to control the cost of informative data is important.*

When there is an appropriately developed model of human performance, running the model on specific tasks of interest may provide useful estimates of human behavior. Using data that is the output from a model of performance is discussed in Appendix B.

This report particularly addresses data collected to measure how well a device supports attention, awareness, and understanding. Several data-types may be helpful for one evaluation issue. Each data-type may be helpful at multiple phases and may be applied to multiple tasks, e.g., differing in scope of activities included.

### *5.2.2. Inspection-based Data*

Observations from inspection are the information gathered when an inspector (who is a relevant expert) examines a device or a representation of the device. By inspection we mean looking at or reviewing something, not using it or observing someone using it. The data resulting from inspection are typically a set of issues or concerns that the expert has identified, perhaps with strengths or overall ratings as well. Different types of materials might be inspected and become available at different phases of development. Once design documents describing the device are available, the design review phase of evaluation can begin. Review of these documents can provide early detection and possible remediation of problems. At very early stages of development, inspection of documents is the primary if not only evaluation data that can be gathered. Later, a prototype or the device itself may be available for inspection. For example, an aviation expert might review a display intended to support awareness of vertical profile to see whether all the information the expert judged important was included in the display.

Inspection data typically provide a list or assessment of problems and thus may be directly relevant to guiding redesign. Various structured inspection methods have been developed in an effort to increase consistency and value of results. However, the inspection results can sometimes be difficult to combine or interpret: data from multiple inspectors is valuable but may be hard to integrate or may even be inconsistent. The value of inspection data depends in large degree on the expertise of the inspector. Expertise is needed to set up the inspection, in terms of what configurations, display, or designed activity flow is to be inspected, and to identify problems in the specific information inspected. Further, some aspects of performance can be very difficult to project (or mentally simulate the operations) just from reviewing, without actually using a system. For example, even experts dramatically underestimate what people will attend to and are aware of. Topics concerning attention, awareness, and understanding may be particularly difficult to assess from many inspection approaches.

A key benefit of inspection data is that it is available early and this can provide a foundation for tracking and identifying problems throughout design and development. It provides a vital source of information identifying (many of the) issues that warrant investigation in performance-based methods.

### *5.2.3. Performance-based Data: Overview and Direct Measures*

#### *5.2.3.1. Characteristics*

Performance-based data provide the best evidence about whether and how a suspected issue impacts task performance using a device. Performance-based measures derive from a person carrying out a goal-directed task. Critically, some aspect of the device being evaluated has to be well enough developed so that an operator can use it in some way. Because the task has a goal, the operator's behavior can be scored as correct vs incorrect or better vs worse. Timing of the behavior from an initial event to the start or to the completion of the behavior may also be recorded, to measure the response speed or the completion time.

Attention, awareness, and understanding affect performance on many tasks, but measuring how well a device supports these more subtle processes is not as direct as measuring a concrete factor such as reachability. Further, measures of a given factor, such as awareness, differ in how direct versus indirect they are. For example, asking a pilot to report the value displayed on a dynamic indicator is a quite direct measure of awareness of that variable, while eye tracking data about where the pilot looked is a quite indirect indicator of awareness.

We discuss two direct measures of attention, awareness, and understanding, one verbal and one nonverbal. These are widely important for measuring performance and particularly relevant to assessing awareness and understanding. We then touch on three indirect measures.

#### *5.2.3.2. Direct Measure: Verbal Report "In the Moment"*

Verbal report uses what an operator (here, the pilot) says as the measure of awareness. In general, if you want to know what someone is aware of, asking them is a good way to find out. Verbal report "in the moment" is an important direct measure. We use verbal report as shorthand for "verbal report in the moment." Verbal report may be a "talk aloud" protocol, narrating what the person is thinking. Alternatively, the operator may be probed to report specific information. Verbal report is a strong measure for assessing the information in awareness or assessing information that can quickly be summoned into awareness, as from a situation model or interpretive frame. People are able to report the content of their model of the situation, either with prompts for specific aspects or as free report,

though report of complex information may be incomplete. Verbal report is used in common methods of measuring situation awareness, including SPAM (Situation Present Assessment Method; Durso & Gronlund, S.D, 1999; Loft et al., 2015) and SAGAT (Situation Awareness Global Assessment Technique; Endsley, 1995).

Verbal report is frequently used to measure awareness of some specific variable, either identified by a probe at the time of report or by instruction at the start of a task. Sometimes this can also measure more complex understanding, as when a correct prediction draws on reasoning from the pilot's situation model. Here, asking the pilot to make inferences based on what they know can be an effective way of using verbal report. Asking "what will happen next" if no pilot actions are taken is one type of inference, which is used in SAGAT. Additional questions can probe the accuracy of the situation model, such as asking what the effect of different inputs will be or the behavior of a mode in the situation.

"Talk aloud" data is related to verbal report but designed to capture the flow of thinking rather than "snap shots" of values of particular variables. Talk aloud is often used to measure problem solving and consists of asking a participant, such as the pilot, to report what they are thinking about moment by moment. Where this has been used, talk-aloud measures do not seem to substantially change how the task is done (after brief, initial practice verbalizing) because people are just saying out loud what they are thinking (Ericsson & Simon, 1984). If the task itself includes talking, e.g. response to air traffic control (ATC), this may fit in as part of the talk-aloud method, but care should be taken that these don't conflict. Talk-aloud measures can be very helpful in assessing how well the pilot understands and can reason about the current situation.

Verbal report "in the moment" is quite different from retrospective verbal report given after the fact; significant forgetting can occur even over short intervals, particularly if attention has shifted or another task has been addressed over the interval. Retrospective report is a particularly bad measure of awareness as people may forget what they were aware of when and reconstruct or guess information. Retrospective report is an indirect performance measure, discussed in that section.

As evaluation data, verbal report has the valuable characteristics of directness, flexibility, richness, simplicity, and lack of interfering with performance.

- *Directness.* Reporting is quite directly linked to what is in awareness at the time of reporting. When a single piece of information is reported, this places very little burden on memory and little room for forgetting. Further, the process of talking is highly over-learned and does not require tapping into any specialized knowledge or difficult-to-execute response.
- *Flexibility.* Verbal report can be used to assess awareness of any meaningful variable. The information to report can be very constrained (what is your current air speed) or quite open (what will happen next; what will you do if). Response options do not need to be specified in advance (as would be the case if the response was instrumented as in button presses) and thus can capture unexpected information. In turn, verbal report can be used to collect open-ended responses, comments, or recommendations as well as very specific, constrained observations.
- *Richness.* Flexibility allows responses to be quite rich. A response may identify the nature of any problems, confusions, or errors and scoring need not just be as correct or incorrect.
- *Simplicity.* Little or no special instrumentation is needed. It may, however, be important to record verbal reporting in some situations, and not just rely on scoring as the participant

talks. If time of response is desired, then software for capturing voice onset may be important. If timing is desired, it can be helpful to include supplementary, instrumented responses, specifically nonverbal intentional, discriminative behavior such as button presses or actions on the interface.

- *Limited impact on performance.* Providing a verbal report of the contents of awareness in response to a probe has relatively little impact on other (nonverbal) task performance. The limited impact is likely due to the highly learned linkage between thoughts and reporting them in speech. Nevertheless, there can be operational situations where responding to a probe question to report a value, or where remembering in advance that all values of a certain type should be reported may be difficult or disrupt performance. Conversely, it is often possible to embed verbal report in the context of an operational task, such as communicating with crew or ATC.

In summary, verbal report is a widely useful type of evaluation data and is particularly helpful for issues concerning awareness.

#### *5.2.3.3. Direct Measure: Non-verbal Intentional Discriminative Actions*

Non-verbal intentional discriminative actions (IDAs) can also be used to measure awareness. To be intentional, the action should be under conscious control, not an action that might be automatic, such as a blink. To be discriminative, the action should be reliably produced to a specific, known class of events and should not be produced without that event class. There are two broad types of IDAs, test-specific responding or responding based on prior-knowledge.

Test-specific IDAs are actions specifically set up for an evaluation, such as pressing a button in response to particular information: e.g., if current altitude is higher than this value, press this button. Test-specific responses may be particularly useful in early testing of a specific issue; to test how well a display supports awareness of altitude, the user might be asked to press one of two buttons whenever the altitude deviated more than 100 feet high or low. Simple responses can be learned to a small class of events to provide a reliable measure of event awareness.

Alternatively, prior-knowledge IDAs are actions learned outside the evaluation context; e.g., a pilot taking manual control of the throttle after becoming aware that the autothrottle disengaged. In certain contexts, taking manual control is a good indicator of awareness that the autothrottle disengaged. Action sequences are often more diagnostic than a single action if the single action might mean different things in different contexts; a sequence of actions may be more constrained, very diagnostic of the person's awareness and understanding, and thus a good IDA. The prior-knowledge type of IDA is a familiar type in studies of situation awareness and is called by several names ("testable responses," Pritchett & Hansman, 2000; "implicit measures" or "implicit performance," Durso & Gronlund, 1999). Knowledge-based responses may be particularly valuable in simulation-based testing where representative pilots serve as participants, operationally relevant scenarios are used, and actions with strong, established association to target information can be used as indicators of pilot awareness.

Assessing a pilot's broader understanding of the situation, or their situation model is valuable, as well as assessing whether they are aware of an isolated value. IDAs are useful here as well. In many situations, carrying out an appropriate response carries strong implications for how a pilot understood and assessed a situation. Simple tasks have been used in experiments to assess how well a pilot integrates information about the current situation with their long-term understanding of how

the airplane behaves (Doane, Sohn, & Jodlowski, 2004) and this type of prediction task could be adapted for use in device assessment to test how well the device supports this type of sensemaking and reasoning. In addition, supporting correct response in the operational setting is the ultimate goal, and so assessing integrated performance is important. The value of having an accurate situation model (based on the currently relevant frame or mental model) is to enable appropriate action.

Thus, IDAs—in addition to verbal report—can be diagnostic measures of awareness of some variables but need to be carefully designed so the actions do in fact reflect knowledge of the variable being assessed. Of course the ultimate goal of awareness of information is for that to support the correct response in the operational situation. Ultimately, attention, awareness, and understanding enable appropriate actions so assessment in an operational context is important.

#### *5.2.3.4. Challenges in Using Direct Performance Measures*

Two major challenges impact use of direct performance measures: 1) guessing, which overestimates awareness; and 2) lack of knowledge about the appropriate action, which underestimates awareness. In using direct performance measures, and particularly action measures rather than verbal report, care must be taken so that the user is very unlikely to be able to guess the correct answer. If a situation is familiar and predictable to an operator or pilot, when probed for the value of a variable they may be able to give the correct response based on guessing. If the issue investigated concerns how well a pilot understands a particular type of situation, a "good guess" or inference may be of interest. However, if the issue concerns evaluating how effectively the flight deck conveys information, then guessing should be excluded. A correct guess does not tell you much specific about the interface. The primary way of reducing guessing is to use scenarios that have unusual elements and thus are hard to predict. In addition, the target response might be scored for greater precision than could be reasonably guessed, e.g. asking for a rapidly changing altitude on a steep descent.

The second challenge is ensuring that the participant does indeed know what should be done on becoming aware of the target information. If a participant becomes aware of the target information, but does not know what to do, this may be incorrectly scored as lack of awareness. It is quite possible to assume too much knowledge on the part of participants. An extensive, carefully designed study that focused on awareness of mode transitions followed up the awareness assessment with questions to assess knowledge (Mumaw et al., 2000). In a noteworthy number of cases, participants did not know the targeted behavior that should be done; in these cases, failure to respond appropriately might have or might not have been due to failure to notice.

### *5.2.4. Performance-based Data: Indirect Measures*

Indirect performance data are data that correlate with the evaluation concept of interest but do not directly measure it. There are several types of data that correlate with how well a device supports attention, awareness, or understanding. These can provide helpful supplementary information, but their limitations should be understood.

#### *5.2.4.1. Retrospective Report*

In retrospective report, participants (such as pilots) are asked, after the fact, about what events they noticed or whether and when they noticed some target event. For example, in a post-simulator training event, a pilot might be asked what the maximum roll had been, or the mode of the autoflight system at top of decent.

Retrospective report is a less accurate measure of awareness, and if measures of awareness are desired, every effort should be made to measure awareness-in-the-moment and not after the fact. While retrospective reporting can be a useful way of assessing a participant's understanding, it is a poor method for diagnosing how well the interface supports awareness. On the one hand, people can forget information rapidly, especially in a dynamic, operational setting; this leads to underestimating how well the interface supported the participant in becoming aware of particular variables or states. On the other hand, people can reconstruct or infer what they believe they were aware of, which can lead to overestimation of how well the interface supported awareness. A retrospective report technique is a better measure of what information can be retrieved from a situation model when asked, rather than an index of state awareness at some time in the past. Meaningful processing of information produces a more stable, retrievable representation; memory for past events (correct or otherwise) may be of interest per se but should not be confused with awareness of current state.

Within human factors, retrospective reports are closely related to methods for collecting subjective ratings of an interface. Retrospective reports of global feelings of awareness are correlated with awareness-in-the-moment and some scales intended to measure situation awareness use retrospective report. The most widely used subjective report scale for the user is SART, Situational Awareness Rating Technique, followed by SA-SWORD, Situation Awareness–Subjective Workload Dominance (reviewed and cited in Jones, 2000). However, user ratings should not be confused with whether and when a user becomes aware of the displayed value of a variable. Relatively recent work includes comparative evaluation of contributions of different measures (Loft et al., 2015) and measure reviews (Durso & Sethumadhavan, 2008; Pina, Donmez, & Cummings, 2008).

Retrospective report has limitations for measuring what users did or did not notice, or the effectiveness of the interface in supporting awareness. Retrospective report of operators will also be limited both because the lack of awareness of an operator is hard to observe and by operator forgetting, even over short intervals. Further, events may be reinterpreted in light of what happened later.

Despite these limitations, it may be useful to collect retrospective reports when "in the moment" measures are not feasible, or to include retrospective report to allow a wider scope of questions. Retrospective reports may be cued by playing a video or verbal prompting, such as asking a pilot why they took a particular action. Such cueing may increase memory for what the pilot was thinking but is subject to effects of later reinterpretation or elaboration. Retrospective report may produce useful information, particularly about how the operator understands the situation, but its limitations should be understood.

#### *5.2.4.2. Eye-tracking*

With the increased availability of inexpensive and unobtrusive eye tracking equipment, measurement of eye tracking and particularly eye-fixation has increased. Looking at, or fixating, a region is, indeed, associated with awareness of information displayed at this region. However, this association is far from perfect.

Information even in the area of fixation often does not enter awareness if the information is unexpected or unrelated to the person's goal. These attentional limitations, called inattentional blindness, change blindness, and inattentional deafness are remarkably widespread (Mack & Rock, 1998; Simons & Rensink, 2005) even among experts. Conversely, people can also become aware of visual information outside the area of fixation. Many factors affect the degree of relation between fixation and awareness, and how these factors play out is not well understood. Thus, if the goal is to

assess whether a participant is aware of specific information displayed at a specific location and time, eye fixations are a surprisingly complex and indirect measure.

Eye-tracking can provide more general information about general trends, such as how gaze, and on average visual attention, is distributed across an interface and how the distribution changes. If the goal is to compare which of two indications is more likely to attract attention following an event, it may be helpful to compare the number and duration of fixations to one region versus the other. Eye tracking and its relation to attention is discussed in greater detail in Appendix C.

#### *5.2.4.3. Physiological Measures*

Physiological measures can provide a useful complement to performance-based measures. They may be generally more helpful in assessing the person's overall response, such as workload or stress, than in assessing how well an interface supports attention and understanding. Again, it may be of interest to assess whether working with a device produces more or less stress than with another, but this is a different assessment issue than investigating how well the device supports the pilot's awareness or understanding. These are active topics of research and relevance of physiological measures to device evaluation are summarized in Appendix D.

### 5.3. Tasks: What Tasks are Used to Produce the Data?

The role of tasks is important though somewhat different for inspection-based methods and performance-based methods. In an inspection-based method an inspector reviews or "mentally simulates" how a task would be done, for example, comparing the current with cleared altitude, and makes some judgement about how accurate, fast, or easy this will be. In a performance-based method an operator (typically a pilot) performs a task and this is measured, typically for accuracy or speed.

Appropriate test tasks will differ depending on the particular device, the issues of concern, and the evaluation phase. It is often useful for testing to move from using small, simple tasks to larger, operational tasks.

Simple tasks can be used to rapidly gather information about a specific set of issues or functionality of the device. A small, simple test task represents an important component of larger operational tasks. Typically, simpler tasks are more easily controlled; useful data may be collected quickly and at lower cost; and simple tasks can be used early in development even when not all aspects of a device have been designed. If a device does not support simple tasks well it is unlikely to fare well in still more complex, operational situations. Simple tasks can help focus later, more complex and expensive assessment on the most relevant operational tasks. Thus, early screening with simple tasks can be very efficient.

Operationally realistic tasks are usually larger scale: they take longer to unfold over time, they occur in a more complex context, and depend on more complex information and actions. Operationally realistic tasks are needed to assess context effects: for many assessment issues performance may be influenced by context, so that direct generalization from a simple to complex context would be unwise. For example, diagnosing the nature of a problem can be much more difficult in a complex, realistic task and environment. Operationally realistic tasks are also important for assessing workload. It may not be possible to collect performance data for operational tasks until late in development and it may also be very hard to judge from inspection how the system will perform across a good sample of operational situations, so performance data here is very important.

Whether for simple or complex tasks, almost always there are more possible combinations of tasks and situations than can be tested. Systematic sampling of tasks and tracking of choices is valuable. Tasks and situations should be sampled to identify and reduce risk. It is valuable to consider including both normal and non-normal scenarios and including familiar and unfamiliar scenarios. For a single "task" of interpreting weather information from screen printouts there are still a very large number of situations that might be presented; an unfamiliar situation might display an unusual weather state; non-normal and unfamiliar situations might be cases where the airplane's ability to fly around weather is impaired or where information is missing from the display.

Where operationally realistic tasks are to be used, there is a very large set of both tasks and situations to select from. Performance testing should consider the broad range of operational tasks and situations that the device might influence. One source of operational tasks is the airline procedure manual. This can be a useful way of identifying routine tasks, both normal and non-normal. However, this will miss important activities that have not been proceduralized, particularly the less routine and less understood tasks. These unexpected, unfamiliar tasks may provide the greatest challenges to attention, awareness, and understanding and so are particularly useful to include in tests. Another listing of tasks relevant to the device is often developed in the requirements and design documents. Many design projects, at least of larger scale, include an analysis or list of the tasks that will use the device. Existing task lists may be expanded, particularly with non-routine and with non-normal tasks (and situations) that are relevant to the device or issue to be tested. A third, more selective source of tasks (and situations) to use can be formed as the output of an early inspection method, Issue-oriented Cognitive walk-through, which we describe below. Results from this method can be organized in a Scenario X Issue Matrix and used to prioritize tasks and situations to use for assessing particular issues.

The selection of tasks or task-situation combinations will differ depending on the scope and purpose of the test. The key objective is to sample for diversity, while still keeping relevant to the test objective. Selecting from a broad list is a great help for converging on an informative, feasible set of tasks. If no task list exists, a task list relevant to the particular test goal can be constructed in combination with scenario design.

### 5.4. Situations: What Situations and Tasks are Used in the Test?

Scenarios specify the situations and the tasks that are to be carried out in that situation. Certain tasks can only be done in certain types of situations and the details of a situation shape what must be done to successfully accomplish a task. While task lists are often available, the useful test situations will likely need to be systematically developed. Whether for early prototyping or late simulator tests, the basic principles for designing a set of scenarios remains the same.

- Scenarios should certainly include normal uses relevant to the intended function of the device, and a sampling across the situations likely to be encountered. The range of situations may be very large, so a systematic way of sampling from broad categories of situations may be helpful. For example, situations for normal and edge cases might categorize types of risk impacts from terrain, weather, airspace, ATC, aircraft trajectory and state, and crew.
- Scenarios should focus substantially on edge cases and expected difficulties since these are likely the most important to identify and reduce risk. One source of difficulty comes from the device itself. For example, if display of certain values of one variable are suspected of being easily confused with particular values of another variable, test trials should be weighted toward these values. The second source of difficulty comes from the flight

situation. Even in early prototyping, some issues may benefit from assessment in situations designed to be confusing or difficult, for example, by imposing a secondary task.

- In addition, scenario design needs to consider how the scenarios affect what the test is measuring. For example, using a certain type of sequence may be appealing because it seems realistic, but it may make events very predictable. In turn, this can undermine assessing whether a participant is actually getting information from a display or is providing a good guess based on prior knowledge and this predictability.

Guidance for developing scenarios with particular attention to the situations is provided in Mumaw, Billman, et al., (2019). Just as issues and scenarios identified in an earlier Inspection-based assessment may help in later performance-based phases, materials developed for prototype-based testing may be helpful for design of simulator tests.

### 5.5. People: Who Provides the Data?

There are three important roles for people in device evaluation; two roles provide data.

1. Inspectors review materials such as device design documents but do not do tasks; inspectors are relevant experts. They provide data from their inspections.
2. Participants carry out tasks; line-pilots are valuable participants. Participants provide performance data.
3. Observers watch, record, or score participants doing tasks; observers may be human factors experts or trained flight instructors and may be the evaluators designing and conducting the test.

These roles are not absolute, but aid understanding different types of methods and tasks.

Selection of appropriate people providing the data is critical for an informative assessment. For inspection-based methods, the inspector needs relevant familiarity with the domain and needs an understanding of the issues that are the focus of evaluation. In addition, the inspector will be much more effective if they are not closely associated with the design and development of the device. Ongoing exposure can normalize problems and reduce the chance of problem recognition. A strong culture of internal critique and design review is very helpful, but it is particularly valuable to bring in "fresh eyes" to look for ways that the device may fail to effectively support the intended function, particularly for any aspects likely to impact safety.

For performance-based methods late in development, it is very important to use participants who are representative of the target users of the device. Test pilots—necessarily—are not representative users. Test pilots may have an exceptional skill level that enables them to compensate for design issues leading to unsafe performance by more representative pilots. This is one of many reasons that all problems should be identified and addressed prior to flight test. For performance-based methods early in development, pilot participants who are representative users are also the best choice. They are most likely to reveal relevant problems and their successful operation is most likely to predict safety in actual operational settings. Further, early, simple tasks, may be not take up much of the participants' time, and it also may be possible to "take the task to the user," e.g., if using static images or a simple desk top simulation. However, the value of including performance tasks early in the design can be so high that they are worth doing even if representative users cannot be used. Some tasks may not depend on extensive piloting knowledge and even nonpilots may be informative participants for simple tasks.

For performance-based methods, the number of participants is an important part of defining the criteria of success. Twelve of twelve successful users means something very different from one of one. Sometimes, the same issue can be tested many times with each user, such as comparing speed and accuracy of navigating through a menu, using two different designs. Issues of practice need to be controlled, but this "reuse" of participants, or "within-subject design" can provide considerable power. Sometimes an issue can only be tested once per pilot or per crew; for example, if recognizing and recovering from a type of surprising event is the issue, performance on repeated presentations may be very different than when the situation arises unexpectedly.

### 5.6. Criteria: How Good is Good Enough?

Evaluation criteria are different for formative versus summative evaluation. Results from formative, inspection-based methods are unlikely to use a success critierion based on a performance standard. However, inspection data may be used comparatively to decide which of alternative designs should be developed further. Similarly, performance data from prototype- and smulation-based methods may be used to decide design and development priorities. It is a matter of negotiation between the FAA and the Applicant what data will be acceptable for summative evaluation. For example, if data from a well-designed prototype test provided compelling evidence that an earlier concern about display interpretation was resolved, this could be used to determine that the device met the requirements of a relevant rule.

Appropriate criteria differ depending on the device and the performance issue. Criteria may be in terms of meeting an independently specified level of performance or in terms of performance relative to another system. For safety critical tasks, such as ability to detect and respond to critical threats, it may be important to determine that all participants were able to accomplish the task, do so within a time limit, and in a manner that does not have undesired effects. For performance that is less immediately linked to safety, such as magnitude and frequency of altitude departures from target value, a more nuanced criteria may be desirable. When time of response is important as a criterion in addition to task success, reference to a maximum rather than average time may be more useful. For example, specifying that 95% of pilots must respond appropriately (e.g., notice and report when a mode becomes inappropriate) within 120 seconds may be much more relevant than specifying that the average response time must be no more than 60 seconds. Criteria may also be specified in terms of acceptably low impact on other tasks. For example, difficulty in finding needed information might delay or disrupt other tasks, or impact workload and task management. If comparison is made to performance in an existing system, how the comparison is to be made should be specified exactly.

The number of participants included in a performance-based evaluation is an important part of specifying success criteria, whether comparing performance to another system or to a fixed-value criterion: the implications of 100% of 3 participants responding appropriately is very different than 100% of 30 participants. Power analysis provides precise methods for determining sample size, given assumptions, but it is not clear how useful this approach is in the evaluation context. Careful analysis and discussion of sample size has been developed for usability testing and the logic outlined there could be applied to safety critical domains (Sauro & Lewis, 2012). Whatever method is used to determine sample size this should be agreed to in advance of the particular test. Where accurate information about a key issue is deemed very important, it may be possible to do a focused, short test on this issue with a (relatively) larger sample size.

### 5.7. Overall Design Considerations and Efficiency

It is desirable if the overall process of designing, evaluating, and certifying a device can be done as efficiently as possible. We have emphasized the value of identifying problems early in design, so that costs of addressing the problems are minimized. In addition, the evaluation process itself can be made more efficient when issues are identified early.

Typically, there are multiple issues of concern as well as multiple rules relevant to certification. Addressing issues and rules one-by-one may be simple and may be efficient for early evaluations. However, overall, addressing issues individually may lead to an inefficient and wasteful evaluation plan, as well as making it hard to detect emergent problems or interactions between issues. For more advanced prototype testing or simulation testing, it is likely possible to integrate assessment of multiple issues not only within the same overall test, but even with a single task. If performance is poor in the task, follow-up might be needed to determine the cause of the problem but if the expectation of problems is fairly low, this testing strategy may be efficient. Combining individual and simultaneous testing of multiple issues is increasingly important for efficiency as the complexity, novelty, and integration of the device increases.

Because of the great diversity of devices to be certified and of the risks and mitigations they pose, there is no general “best” or “cookie cutter” evaluation. Rather, certification and the evaluation standing behind that certification must be individually tailored. Broadly, some performance data are better than none, so long as the limitations of the data are borne in mind. Understanding the method choices and their merits enables specification of a useful plan for a specific case. Each also requires FAA judgment of device risk and evaluation costs.

Planning and, when called for, re-planning are important for efficiency. Effective planning is a key factor for getting the most relevant, diagnostic information for effort spent. Identifying issues early is a key enabler, as well as integrating testing of multiple issues. Ideally, the plan through evaluation to successful certification can be specified in advance, allowing highly efficient use of resources. However, information gathered early will change the understanding of risks and strengths associated with the design and make adaptation of the plan desirable. Attention to planning and re-planning is likely to be time well spent.

Finally, while flight testing is a critical check that the as-built aircraft does fly as designed, it is the least effective and least efficient form of assessing the design: it occurs very late; it is very expensive; and it can test only a very limited combination of tasks and situations.

## 6. Evaluation Methods for Design Review Phase

### 6.1. Purpose and Scope

A key benefit of evaluation based on review of design materials is that it can be carried out early in development when addressing an identified problem is lowest cost. Design review methods primarily help form the design rather than provide a summative “seal of approval.” Nevertheless, the design can also be inspected for clear violations of some rules. Where human factors issues are well understood they may be addressed by standardized rules or by Applicant standards. The design may be inspected for violations of such rules. For example, threats to awareness from violations of rules defining the colors required for different types of alerts could be identified from a well-specified design statement.

As with performance-based methods, scope of design review methods can be adjusted to fit within available resources. Similarly, the focus of review can be adjusted to address the most novel, interactive, or otherwise most risky aspects of the device. Breadth and depth can be modified over the course of the inspection.

We describe five systematic, inspection-based methods applicable to evaluation of support for attention, awareness, and understanding. Cognitive walk-through, Heuristic Evaluation, and Cognitive Performance Indicators are quite general methods, while Information Requirements Analysis and Human-computer Interaction Process Analysis focus more specifically on supporting attention, awareness, and understanding,

We begin by describing Cognitive Walk-through and a novel adaptation of this method we tailored to fit evaluation of device-support for attention, awareness, and understanding. This adaptation: a) emphasizes inspection of scenarios, that is, combinations of task and situation not just tasks; and b) co-ordinates inspection by task and inspection by issue. Next we turn to the Heuristic Evaluation and Cognitive Performance Indicators methods, which can provide helpful descriptions of issues to consider. Finally, we consider the benefits of the narrower, more focused methods.

## 6.2. Cognitive Walk-through and the Issue-based Cognitive Walk-through Modification

The traditional Cognitive Walk-through method begins with a task analysis to identify all the tasks for which the device, program, or interface will be used. A team steps through the process of carrying out the selected set of tasks, and at each step looks for problems. Key elements for effective use include specifying a relevant and sufficiently thorough collection of tasks and the situations in which the tasks are done, as well as ensuring that the evaluators themselves know the right way of doing the tasks (Polson, Lewis, Rieman, & Wharton, 1992). This method was developed with a focus on software intended to be "walk-up and use" but can be adapted to technical applications such as the flight deck.

Cognitive Walk-throughs can be done with varying degrees of coverage of tasks and issues. Adaptation to use for flight-deck evaluation will benefit from strategic rather than exhaustive coverage. We suggest a novel modification to make the method more helpful for use in flight-deck evaluation: 1) selective inclusion of scenarios and issues; and 2) dual organization by task-situation pairs (scenarios) and by issues. The combination of scenarios and issues can frame the scope and focus of the walk-through, so these should be selected with the purpose of this particular walk-through in mind. Here are the steps for applying this method of cognitive walk-through:

1. Concerning tasks, the evaluator identifies a set of task-situation pairs (scenarios) that will be focal for the walk-through. Scenarios should include:
    a. focus on tasks and situation central to the intended function of the device;
    b. scenarios that may be affected by the device as "unexpected consequences;
    c. scenarios that may affect ability of the device to operate as intended operation.

    Airline procedures can form an important source for tasks, but consideration should also be given to activities that have not been proceduralized, such as operating in normal but unusual combinations of conditions or in non-normal multi-fault situations. These should be listed with a description of the task-scenario pair and a name. If there is a natural order in which the tasks are typically performed it is helpful to list tasks in that order.

2. Concerning issues, the evaluator needs a set of issues or topics which characterize possible problems with the support provided by the device, which are to be checked. The evaluator can select (and extend) issues as in our example in Appendix E or from the longer list presented in (Mumaw, Haworth, Billman, & Feary, 2019). If there is a particular area where more specific concerns are suspected, the issues can be broken down into finer categories. Conversely, if a rough check for more peripheral topics is desired, just a high-level characterization might be used.
3. Once the tasks and issues to be inspected are identified, a matrix listing these is built as illustrated in Table 3 and Appendix E. A summary column is included for gathering observations about each issue across tasks. An annotation row is added for concerns about a task that did not fall easily into the issues being inspected. This provides a matrix for making and recording observations to identify problems. Having easily expandable cells for notes and observations is important. A schematic example is shown in Table 3 for issues concerning attention, awareness, and understanding. The matrix may also have some annotation space for issues that are not linked to a particular task.
4. The matrix provides both a structure for conducting the inspection and for recording the observations. The inspector steps through each scenario and issue combination. A scenario-organized inspection selects a scenario and steps through its issues, looking for problems related to each issue. An issue-organized inspection selects an issue and reviews that issue across scenarios. It may even be efficient to scan first checking each issue to see which scenarios are most impacted by that issue and then review by scanning by scenario to see which issues most affect each scenario. Taking first one and then the other perspective provides an informal cross-check. Issues not noticed initially may become apparent after reviewing multiple scenarios. Once all the scenarios have been reviewed for an issue, the observations for that issue can be summarized in the Issue Summary Column. This can also provide an efficient way of comparing results from more than one inspector.

We suggest this as the basic method of application, but there are multiple variations. For example, it may be useful to build several smaller matrices, each focused on a more limited set of issues. If needed, this type of structured inspection can be carried out at a coarser or at a more fine-grained level or done at a coarse level first with problem areas then inspected more closely. It is extremely valuable to have more than one inspector, ideally with somewhat different backgrounds or experience; there often is a big gain in problem discovery from three inspectors. Nevertheless, observations from a single inspector can be very valuable.

| Table 3. Matrix for Specifying and Tracking the Potential Issues, the Defined Situations, and Result from an Issue-oriented Cognitive Walk-through | | | | | |
|---|---|---|---|---|---|
| *Scenarios (columns)* Issues (rows) | *...* | *Task & Situation: Scenario I* | *Task & Scenario J* | *Task & Scenario K* | *Issue Summary Column* |
| Needed information is missing. | | | | | |
| Perceptual representation of needed information is inadequate. | | | | | |
| Labels, icons, or messages may be hard to identify or understand. | | | | | |
| ... | | | | | |
| Interface does not support projecting future events. | | | | | |
| The interface does not support accessing or weighting cost/benefit of relevant alternative action choices. | | | | | |
| ... | | | | | |
| Comment row for scenarios. | | | | | |

*Note: A scenario is a combination of a task or set of tasks in a particular situation. An issue is a topic of concern for the evaluation. The table shows a sample of issues and scenarios with existence of additional rows and columns indicated with the elision "..." punctuation.*

### 6.3. Heuristic Evaluation

The Heuristic Evaluation method provides the inspector with a list of quite general characteristics that may impact usability of a system, such as error prevention, error recovery, and reliance on recognition rather than recall. Nielsen (1994) provided a set of 10 heuristic tests and several other categories have been suggested as well. For complex technical systems heuristic evaluation can provide useful insights but may not be the most effective method. Nevertheless, these heuristics of design are related to topics in our "Issues" list and may be useful to keep in mind in any evaluation.

1. Visibility of system status
2. Match between system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetic and minimalist design
9. Help users recognize, diagnose, and recover from errors
10. Help and documentation

A useful overview is provided on Wikipedia and more detailed information is available (https://en.wikipedia.org /wiki /Heuristic_evaluation and in Nielsen, 1994).

### 6.4. Cognitive Performance Indicators

A set of Cognitive Performance Indicators and an associated analysis tool have recently been developed to help assess complex, technical software (Wiggins, Cox, & Patterson, 2010). This is developed from within the Cognitive Engineering and Naturalistic Decision-making evaluation traditions. This approach is similar to heuristic evaluation, but the characteristics to be checked are somewhat more complex, and evaluation may be by feature as well as by task. The heuristics topics for assessment are:

1. Option workability
2. Cue prominence
3. Direct comprehension
4. Fine distinctions
5. Transparency
6. Enabling anticipation
7. Historic information
8. Situation assessment
9. Directability
10. Flexibility in procedures
11. Adjustable settings

This approach seems quite relevant to aviation, but likely requires more training or familiarization with the approach than does Cognitive Walk-through as the characteristics are more abstract. Nevertheless, considering these dimensions may be useful to include in an inspection and may lead to noticing issues not flagged by a more traditional set of heuristics. Notice that again there is overlap between the cognitive performance indicator heuristics and issues listed in (Mumaw, Haworth, et al., 2019).

### 6.5. Information Requirements Analysis

This method draws on an analysis of the work domain to identify what information is needed for work within the domain (Burns & Hajdukiewicz, 2004). The device can then be checked for whether it provides the needed information. For device evaluation, the intended functions of the device provide a description of the work domain. In doing an Information Requirements Analysis it is useful to think quite specifically about the information. For example, the difference between two values might be a variable of interest in its own right, such as distance from stall speed or difference in altitude restrictions of adjacent way points. If a difference or ratio between other variables is important, that derived variable should be included in the information requirements analysis. The design can then be reviewed for how directly needed information is provided versus how much the pilot must calculate the needed values. For example, the pilot might calculate a sufficient distance for required descent using a 3-to-1 rule of thumb; alternatively, a calculator might provide this, or provide a more precise calculation including windspeed.

### 6.6. The Human-computer Interaction Process Analysis

The HCIPA method (Human-computer Interaction Process Analysis; Sherry, Medina, Feary, & Otiker, 2008) and an earlier version RAFIV (Reformulate, Access, Format, Insert, Verify; Sherry,

Feary, Polson, & Fennell, 2003) provides a structured method for assessing how clearly each step in a task is cued. This method focuses on determining the cues needed to identify tasks, the ease of communicating the tasks to and from the systems, and the indicators needed to verify that the tasks are being accomplished.

The method and its application are described in several papers and a spreadsheet tool and an electronic version have been developed (Fennell, Sherry, Roberts, Jr., & Feary, 2006; Medina, Sherry, & Feary, 2010; Sherry et al., 2008). The tool provides a structured method for entering actions needed to complete each of a set of tasks, as illustrated for an air traffic management task (Medina et al., 2010). Key inputs to the tool are judgements of how saliently each action needed to complete the task is cued. The software organizes tasks and integrates judgements to score how saliently cued the needed actions in the task are, overall. This can be used to systematically identify where an interface provides poor and possibly inadequate support for a task. This method focuses on execution of action sequences to accomplish tasks. However, the points where an action is not adequately cued may also identify places where the interface does not clearly provide information needed to understand the situation or to know what action is needed. As with other structured design-review methods, this may both provide guidance for design changes early and for what tasks may be particularly informative to include in performance-based assessments.

### 6.7. Considerations for Design Review Phase

Design review provides two types of benefits:

1. Early problem identification can allow early response and adaptation when the cost of change is very low.
2. Issues identified here can guide later, data-based evaluation, and focus those resources on functions and situations that may pose greatest risk.

Despite the value of early inspection, not all problems may be discovered by inspection. People often have quite inaccurate intuitions about limits on attention and awareness, making performance data particularly valuable for these topics. Without data-driven assessment from later evaluation phases, decision makers may lack a well-motivated basis for assessing safety or making design choices.

## 7. Evaluation Methods for Prototype-based Phase

### 7.1. Purpose and Scope

Prototypes allow an operator to carry out tasks, and thus provide the first opportunity to gather performance data for the evaluation. They enable data-driven assessment of how well the flight deck supports pilot attention, awareness, and understanding. Performance data for these factors is very important for multiple reasons.

- People have attention limits with critical implications for safety. The device needs to support the pilot in maintaining awareness within attentional resources.
- Assessing how well the flight deck supports pilot awareness cannot be reliably estimated based on inspection and intuition. People, even experts, do not have good intuitions about what, and specifically how much, information will enter an operator’s awareness. Rather, people systematically overestimate what will be noticed. Assessment critically requires data that are performance-based and tailored to the issue in question.

- To understand the current situation, awareness of current values of variables needs to be integrated with the pilot's understanding of how the airplane behaves. This enables the pilot to manage risk and to take appropriate action. The fundamental function of the aircraft displays is to support the pilot in updating their understanding, and thus, their action choices.

Evaluation should check for issues with the pilot "pulling in" information to support the current goal, as in monitoring, when attention is driven by endogenous factors. It should also check for issues with directing the pilot's attention to critical if unexpected information, as with response to alarms, when attention should be driven by exogenous factors. In addition, prototypes can be used to evaluate support for prediction, assessment, and understanding, such as making a timely decision that the aircraft is unable to make a restriction.

The purpose of prototype testing may be issue identification or issue assessment, for the issues already identified. Issue-identification tests are likely broader than tests to assess an already-identified set of issues. Issue-assessment tests are likely to focus specifically on activities and situations directly related to the identified issue. It can be very useful to assess a specific issue with an early, simple prototype. Prototype evaluation is often formative and allows comparison of performance using alternative designs. As noted, inspection methods described for design review can also be used when prototypes or simulators are available.

Multiple prototype tests may efficiently complement each other in an overall evaluation plan. Prototype-based evaluation can be low cost. Prototypes can differ in their scope and realism and can be developed to target issues of particular concern. The cost of developing a targeted, modest-scope prototype may be very low, though the product is very useful for a particular issue of concern. For example, a test goal might be assessing how well a new indicator's labels convey the intended information quickly and accurately. Here short test sessions, with few participants using a simple prototype may provide most of the needed information. Higher level tasks can also be assessed even with simple mock-ups, for example, sequences of static images showing weather or navigation displays can be used to test support for pilots interpreting and reasoning with the information presented.

Design a prototype study by considering the:

- issue(s) to be tested or type of issue to be discovered (specific purpose)
- relevant component/indicators in the device
- tasks or activities that use the component
- situations in which the tasks are done
- feasible scope given resources

This section includes information also relevant to simulator-based testing and the information in Section 6. Evaluation Methods for Simulator-based Phase may also be helpful for the prototype-based phase. A top-level consideration of design of both prototype and simulator phases may be helpful.

### 7.2. People: Participants Provide the Data

Participants provide more useful information when they are not associated with development of the system being evaluated. Whether the participants need to be pilots will depend on the nature of the awareness issue being assessed. In the case of simple reporting tasks, nonpilots may have a useful

role. All participants should have an initial familiarization with the displays and their content prior to the test trials.

## 7.3. Stimuli

The stimuli are a series of displays or events that include the target variables in a relevant scenario. The series of displays includes the indicator(s) being evaluated across the range of values of interest, in the contexts relevant to the issue(s). Depending on the issue, stimuli may be static images (particularly for very early evaluation), simple animations, or displays in a part-task simulator. Presentation software can be used to create simple animations. Screen shots of simulators can be used to create context into which a novel component can be edited.

## 7.4. Equipment

The prototype and any additional measurement instruments are the equipment. Computer-based prototypes are very valuable for recording responses and collecting timing data, when available. It may be quite feasible to build a dynamic, computerized prototype using simple presentation software, if the scope is narrowed to address an aspect of key concern, and buttons can be instrumented to record responses. External timers, particularly for longer time intervals can also be helpful if the prototype is not able to record timing data.

## 7.5. Procedures

The appropriate procedure depends on the evaluation goal. The goal might be to assess how well the interface supports directed search where a participant is explicitly looking for the target variable (monitoring) or how well the interface supports "noticing" of a variable when the participant doesn't know in advance that variable will change or need attention (flagging and alerting). Support for "noticing" information may differ among changes that are just unexpected, are surprising, or are startling. The evaluation goal might emphasize accuracy or speed. If accuracy is the primary interest, the conditions should be difficult enough to produce substantial errors, such as with short display time or high workload from multiple tasks. If speed is the primary interest, the display might be presented until the participant responds.

### *7.5.1 Procedures for Assessing Awareness of Specific Variables*

- To measure awareness, the direct, verbal report in-the-moment is a good choice.
- The participant views a display and reports the value of the target variable(s) for a single trial. In prototype-based testing many quick trials by one participant can be used for a test.
- Different variations in the procedure for a trial will change what is being measured:
  - Awareness vs Memory: How much is reported on one trial. If awareness and not memory is the issue, the participant should report only one or a very few variables and do this promptly once the display no longer shows the value.
  - Monitoring vs Alerting: What does the participant know in advance about what to report. If the main interest is how well the display helps the pilot find the information they are looking for, as in monitoring (hence endogenously directed attention), then the participant should know in advance what variable will be reported. They will be looking for what they will be probed for and performance will reflect how well the display provides information they are seeking. If the main interest is how well the display pulls the pilot's attention to particular information, as in alerting (hence exogenously directed attention), then the participant should not be able to predict in advance what they will be

asked; their ability to report the probed value will then be driven by how much the display captured their attention.

### 7.5.2. Procedures for Assessing Broader Understanding or Situation Model

- An evaluator may want to assess how well a device supports a pilot in finding, retaining, and integrating information to make sense of the situation. For example, a new weather display may have the intended function of aiding the pilot understand current and upcoming weather; a new mode control panel may have the intended function of aiding the pilot understand and predict the behavior of the autoflight system.
- Identifying the scope of relevant information to probe is critical.
- Adapt existing methods to testing devices. Several established measures of situation awareness use verbal-report-in the moment (SPAM, SAGAT). Those methods are designed to test a pilot's global awareness, not device effectiveness. These methods can, however, be modified for testing device effectiveness. Specifically, rather than randomly sampling across a global set of variables, the pilot would only be asked about information relevant to the intended function of the device. This might include questions about what the current state is, explanations of what actions produced it, predictions about what will happen without pilot intervention, and what pilot intervention would be a good method to produce a specified outcome.
- If the goal is to assess how quickly the pilot can integrate information, then rapid displays, quick probe of a random question among those relevant, and timing of the response may be useful. If the goal is to assess how elaborated or enduring the pilot's situation model or understanding of the situation is, then intentionally delaying the verbal report may be useful. In addition to reporting the value of variables, the pilot might also be asked to explain the basis for predicting a particular future state. The pilot might be asked "what-if" probes. For example, the pilot might be asked what would happen if a particular mode were selected in the current situation, or to decide whether it would be possible to make a way point if a particular clearance were given in the current situation.

### 7.5.3. Procedure Variations

There are several ways this basic method can be varied, depending on the evaluation goal.

1. Does the user know in advance what to report (the target)? If the target is only specified after the target information is removed, an accurate report can only come from what the user already noticed; this report method measures what the user has been tracking or looking for. This is a good measure of what the user is aware of "on their own," or what information is in their current situation model. If the report-target is specified when the information is still available, this measures the user ability to find the target information combined with whether they already know the target. This can be a good measure of how easily the target information can be found, and can assess issues from whether labels are clear, whether an alert successfully communicates the action needed, or how long it takes to navigate to the target information. Timing information can be particularly useful here (as collected in SPAM Durso & Dattel, 2004).
2. How much context is needed for a useful test? Context may come from the flight deck, the environment, or personnel. Some issues can be addressed with very little context and a "standard" environment. Some issues may be addressed in isolation from other flight deck components, for example assessing whether the intended information is accurately read off from a set of indicators. Other issues may require more of the flight deck (or cockpit) context, for example, assessing whether understanding an indicator in the new component

is confused by similar terms used differently in other parts of the flight deck, or assessing whether the amount of information presented overall makes it hard to detect changes in information displayed in the component being tested. Concerning issues about the physical or task environment, sometimes the relevant conditions can be created to allow an early assessment, such as whether a display can be read at low luminance or whether information will be noticed with a heavy, unrelated workload. However, assessment of support under varied environmental conditions may best wait until simulator testing.

3. Does the user just report a displayed value, or does the report tap into additional knowledge and a more complex task? If the issue concerns clarity of the indicated information, asking for a direct report is the clearest means of assessing this. If the issue concerns support for integrating displayed values into an integrated or action-oriented situation model, questions or actions that include other knowledge can be useful. Informative questions include asking for predictions of aircraft behavior or state or for explanations both of how to produce certain states or why the aircraft is in a certain state. Correct answers here depend on awareness of indicated information, an understanding of how the aircraft works, and ability to integrate information to produce a prediction or explanation.
4. Is speed or accuracy of primary concern? Accuracy of the reported value is always important, and sometimes this is the primary issue being assessed, for example, assessing whether information is likely to be misinterpreted. In some cases, precision of the value may the issue. Speed of response may be of primary concern, or accuracy within an available time. For example, is transient information reliably understood within the time it is displayed?
5. Sometimes the question is how long it may take to access the correct information; for example, how long is the time to navigate, reconfigure, or otherwise manage the flight deck to access needed information.

## 7.6. Test Materials: Tasks, Situations, and Scenarios

Scenarios consist of the task and situation. These are presented to the operator and performance on the task in the situation is measured. The set of scenarios for a test should be designed to provide the most informative sample that is feasible for the prototype, test goal, and available resources. Getting the most value from a test means sampling the situations broadly, with a particular focus on situations likely to be problematic for the device being tested. Typically, there will be more tasks and situations where the device is used than can be included in a test. Rather, strategic sampling of scenarios is needed, which still provides good coverage. Methods for developing scenarios for test materials are detailed in Mumaw, Billman, & Feary (2019) and a method for developing a set of scenarios and tracking how this covers the issues to be tested is summarized in Appendix E.

# 8. Evaluation Methods for Simulator-based Phase

## 8.1. Purpose and Comparison to Prototype-based and Flight Test Evaluation

Ultimately, attention, awareness, and understanding enable appropriate operational actions. Thus, assessment in an operational context to carry out flight tasks is very valuable. Simulator-based testing is the efficient way of doing this.

Relative to prototype-based evaluation, key advantages of a simulator test environment are: 1) realistic dynamics for aircraft and its context; 2) richer control actions available to the pilot; and 3) integrated support that allows extended, varied situations and activities within one complex setting. With respect to attention, awareness, and understanding, these advantages are particularly valuable for assessing interrelated issues concerning:

- riming or timeliness, such as how quickly a pilot becomes aware of an alert
- how operational context impacts basic processes of attention and awareness such as noticing a changed value in an indicator
- how pilots build and update an integrated understanding of the situation

Prior prototype testing can help guide the resources for simulator-based evaluation to where testing will most reduce risk.

Relative to flight testing, simulator-based testing also has large advantages. It can include unsafe conditions not feasible in flight-test and conditions difficult to control. Many more trials and scenarios can be included than in flight testing because of its lower cost in time and money. The human-factors issues should be resolved before simulator-based testing is complete, with no known issues remaining at flight test.

## 8.2. Designing a Simulator Assessment

### *8.2.1. Define Overall Objectives of Testing and Test Strategy*

Frequently, multiple simulator tests can contribute to the overall assessment of the same device, possibly using multiple simulators of different fidelity. Multiple, low fidelity simulator tests may be more effective than one large assessment. Where a large test is planned, it can often be designed to test multiple goals.

As with prototype testing, simulation testing may be intended to identify issues or to assess the effects and severity of known issues. These tests may have different scope. Issue-finding tests may sample a broad set of conditions to scan for issues that were previously missed, emerged late in design, or result from unexpected interaction with other components. Issue assessment may select the situations judged most likely to reveal whether and how an issue affects performance. It may be possible to embed assessing a particular issue or issue set within a broader issue-finding test. The simulator-based test plan may evolve as information is gained but considering the overall resources and the identified risks together can lead to a more efficient testing strategy.

Describe each identified issue to be tested and/or characterize any types of issue particularly relevant to discover. The key goal of designing a test is make sure it provides best information relevant to the goal. The remaining test-design steps need not be done in a strict order and decisions on one aspect may affect others.

### *8.2.2. Identify Performance Measures*

For each issue, identify the performance measures to be used. Measures may go beyond pilot activities in normal flight, such as more detailed reporting or unusual dual tasks demands. Testing may need to balance operational fidelity and informativeness of tasks and measures to provide the most useful information about the issue.

Multiple performance measures are often helpful, such as both verbal report and IDAs. Just as for IDAs, verbal report can be talk that is part of normal cockpit procedures (prior knowledge about communication in the coc pit) or reporting specifically requested for the test (test-specific, e.g. an added call-out). Test-specific measures should not conflict with the required flight operations. Simulation scenarios may be set up to require particular discriminative actions, which can make prior-knowledge IDAs particularly informative.

Measures need to be linked to the specific issues being tested. Below are examples of verbal report and IDAs relevant to testing issues related to attention, awareness, and understanding:

- *Prior-knowledge Verbal Report.* Example: Time to report "unable." Time from provision of an ATC clearance until pilot monitoring (PM) makes statement indicating aircraft is unable to make the clearance. Percent of trials where pilot states unable would be measured as well, but this might not be very informative if the problem were virtually always identified by the PM. This can assess how well the device supports PM awareness of current trajectory and the feasibility of trajectory modifications.
- *Prior-knowledge IDA.* Example: Vertical deviation correction. Time from when log data shows airplane is 100 ft. above clearance to time corrective control action is taken. Verbal report of problem and action could be used as measures as well, as could the type of corrective action taken. Operational measures such as duration of period aircraft is more than 200 ft beyond clearance may be valuable as well, though a more indirect measure of pilot situation assessment.
- *Test specific Verbal Report.* Example: Trajectory prediction. The sim flight deck of the PM is covered and/or the sim may be halted. The PM is asked to state what will happen "next" if no crew action is taken. This is an extension of methods sometimes used informally in early training, where the instructor may cover some of the student's instruments.
- T*est-Specific IDA.* Example: Pilot is given a secondary task to press a button whenever a specific type of event occurs, such as a change in vertical speed greater than some amount. Task-specific actions may be designed in order to add workload where this is desired. When flying, test-specific IDAs probably will increase workload to some extent.
- Prediction and explanation: Pilots can be probed for past, present, or future states and asked to explain the changes.

In addition to measuring attention, awareness, and understanding, a simulator allows collection of data concerning the control and configuration actions taken by the pilot. Specific performance measures can be included that indirectly depend on an understanding of relevant information of interest, such as time past the programmed top-of-descent (TOD) before descent begins.

### 8.2.3. Identify Data to Collect

Different measures require different sorts of data, so data collected needs to support the measure. Typically, data need to be coded or scored. Data might be collected and scored directly by a human rater/evaluator or collected in log files, from a computer prototype, a simulator, or eye tracking equipment. Video provides another data stream from which measures can be coded, reviewed, or anomalies checked after the simulation has run.

Direct coding by a human observer during the simulation runs can be similar to what an instructor pilot would do in a training simulation, but typically requires more structure and selective noting of the behaviors relevant to the testing. This real-time coding can capture a variety of relevant higher-level behaviors and make these immediately available during the assessment. Real-time coding has the disadvantage that an observer's ability to notice the relevant behavior will be limited and will be vulnerable to distractions. An observer should have a limited, well-understood set of events to be recorded and a well-organized logging tool. (Flight test cards serve this function in flight testing, typically for very high-level observations.) The observer's logging tool (such as a spread sheet) can capture entry-time as well—this can be very useful if the timing of interest is at the scale of minutes or longer. Simulators can also be configured so that an event marker can be added to the simulator data. Review of performance after the simulation run adds time but can be a valuable addition to check the real-time coding or to identify key behaviors.

The key to efficient observer-based coding, whether real-time or post-simulator, is: a) specification in advance of exactly what events the observer should be marking; and b) providing the observer with a clear, fast method of recording data. Instructor pilots might be trained to carry out the coding or this might be done by an evaluator on the testing team.

In addition to the low-level data such as control actions and aircraft state, the raw simulator data can be post-processed to derive more complex variables. These might include max heading deviation before correction response, maximum roll in autoflight before manual engagement, or time difference between actual and predicted waypoint crossing. Post-processing can be time-consuming, so may be best limited to address particular issues. In addition, it can be valuable to build a library of post-processing functions, if there will be recurring use. Simulator log files can be extremely valuable for capturing precise timing, such a time between an indicated state changes and a corrective action is taken. If multiple data types are used, it is very valuable to have them all consistently time stamped to allow integration.

### 8.2.4. Define Scenarios: Tasks and Situations

Design scenarios by identifying the task goals and conditions in which the goals should be carried out. Draw on scenarios developed for prototype-phase testing, as appropriate. Ensure scenarios cover both the edge cases and core cases relevant to the device and its intended function.

For issue-driven assessment, focusing on edge cases may be particularly helpful in determining whether issues have been successfully addressed. Edge cases look at scenarios where the device may be at its limits, where poor functionality of the device may have high safety impacts, or with combinations of conditions not envisioned by the designers. Core cases are important as well, to ensure that the device supports its function in normal use. Core scenarios might include general conditions likely to make normal use difficult, such as high workload and rapidly changing, unexpected events, even if not directly related to the device function. Core cases may be particularly

valuable when looking for possible contextual effects. Typically, scenarios should be unfamiliar so that participants cannot excessively anticipant events, particularly those intended to be surprising.

Managing surprise is often important but difficult. When performance concerns a rare event, the relevant performance is when the pilot is not expecting the event. Thus, if a pilot is able to predict a particular type of event in advance, the data may not be very representative of actual performance. Further, for typical events, the pilot may be able to guess current values or what will happen next, thus eliminating actual use of the displays to be evaluated. Setting the pilot to expect one (or more) goal or type of event but testing for another can be one way of managing surprise, and a potential benefit of integrated testing of multiple issues.

Scenarios can be designed as "snippets" or as longer spans of flight. To assess a specific, limited issue, it may be efficient to use snippets that focus specifically on that issue. Many short scenarios can be run quickly, collecting targeted information quickly. If the issue is only relevant in a particular phase of flight or set of conditions, this can be a useful strategy. Events in shorter, more repetitive snippets are likely to become expected; therefore, this may not be a good strategy if the intent is to produce unexpected, surprising, or startling events.

Sometimes, the test sequence of events can be completely programmed into the simulator (or earlier prototype). Other times, it may require human involvement of "confederates" instructed to act in a specific manner. This might fill the role of ATC clearances or behavior of the instructor pilot managing the simulator. A powerful though controversial possibility is to have one of the crew carry out scripted actions to produce the intended situation.

### *8.2.5. People: Participants Provide the Data*

Use pilots who are appropriately trained and who are representative of the customer's pilots as participants. In addition, select for variability within this group and particularly include less experienced pilots, or pilots who are more likely to have difficulties with the device. Just as focusing on edge cases is a way of finding hard-to-spot problems, use of less expert pilots can be an efficient way of revealing problems. Number of participants needed is influenced by success criteria and should be established in advance. Here it may be helpful to be able to run quite a few users in a short test to assess a critical aspect, but run a smaller number for a longer, broader test. The ability to run the simulator at a convenient location may aid inclusion of a larger number of pilots.

### *8.2.6. Criteria and Analysis*

It is valuable to establish at least tentative performance criteria in advance, specified in terms of performance on identified measures. See Section 5.5.6 on criteria. Processing simulator log files is time consuming, typically requires deriving measures of interest, and ideally should be designed in advance of data collection. Time needed for analysis versus precision of measurement is a relevant factor in deciding how much to rely on observer coding versus analysis of log data. Libraries of analysis routines are very useful to build up, whenever there is the possibility of reuse.

### *8.2.7. Assessing Multiple Issues*

Where assessment concerns multiple issues, it may be useful to test them individually or together. While it is important to consider whether testing one issue might interfere with test of another, there can be important benefits from integrated testing. Single-issue "snippets" can be very useful in addressing a particular issue, but they reduce the similarity to actual use. It is often possible to

integrate testing of multiple issues by linking together situations likely to be diagnostic of different issues. Thus, if all the issues and their possible test strategies are considered together, it may be feasible to design a longer and more complex scenario that is diagnostic, more realistic, and more efficient. A key advantage of integrated assessment is greater ability to make events, particularly problems, harder to anticipate and preserve a greater aspect of surprise. The complexity in integrated assessment can make it less easy for the pilot to 'guess what the simulation is about.' For example, if an auto throttle disconnects part way through a long flight, the disconnect can be truly unexpected, thus providing a more relevant assessment of alerting adequacy.

If performance remains good in the conditions designed to assess multiple concerns, this provides evidence that the concerns are likely resolved. However, if there are performance problems it may be difficult to identify what aspect(s) of the device contribute to the problem, and follow-up testing that decomposes the issues may be needed.

## 8.3. Hints and Cautions

### *8.3.1. Simulation Costs and Benefits*

A wide range of simulation tools are available: part-task trainers, simulators of different fidelity levels, and applicant programming to simulate interaction with simple devices may even be feasible. Pick the simplest simulator sufficient for the testing purpose. A simulation of the whole device being evaluated will be available only late in development. Note, however, that if a dynamic model of some limited aspect or component would be very valuable, a simulation of part of the device in isolation may still be useful, and a prototype might address this. Compiling and understanding the results from a complicated simulator test with multiple events of interest may take some time, so identifying the scope and resources in advance is a good idea. Building for reuse of scenarios and of methods can be helpful.

### *8.3.2. Strengths and Weakness of Measures*

Verbal reports can often be framed in naturalistic ways such as the need to communicate with the other pilot. Generally, if the report is about the pilot's activities, reporting does not disrupt those activities; conversely, if a pilot is asked to report information unrelated to those activities, i.e. a secondary task, reporting may add a significant burden and it may be used strategically to increase workload.

IDAs can be particularly valuable in simulator-based testing of issues concerned with awareness. A well-designed IDA avoids under- and over-estimating awareness of the variable it is intended to measure. Ideally, the pilot only and always does the IDA if they are aware of the variable the IDA is measuring. Pilots should know to always take the discriminative action when aware of the relevant variable value; if a pilot is aware of the autoflight mode but does not know that the IDA is the correct response, this will underestimate awareness of the variable. The pilot should only take the action when the target value is known; if the situation is highly predictable so the action is cued by other information, then the pilot can "guess" the action without being aware of the target variable and this will overestimate awareness of the variable.

Eye-tracking and physiological measures may be helpful for identifying when events of interest are detected. It may be useful to know when gaze moves to a particular area of interest. In combination with verbal report this may provide useful information about when information is likely entering awareness. These are discussed in Appendices C and D. For the purposes of evaluating equipment, these measures are very indirect.

### *8.3.3. Simulator Evaluation Issues and Tradeoffs*

Pilot performance on tasks assessing information awareness or evaluation is affected by how well the pilot can predict what variables will be probed. Thus, a response to a first, unexpected probe may give “normal” speed and accuracy, while later, predictable probes for the same variable produce “best possible” rather than normal performance. Data of both types may be useful, but “best” performance should not be confused with “normal” performance.

Simulation-based evaluation should include good methods for determining the full set of scenarios sufficient to test the issues. This will include a method for sampling the combination of normal tasks by normal situations and a method for identifying the challenging, non-normal, and “edge” cases to include.

Designing simulator scenarios to test multiple issues can be valuable because simulation testing is expensive and complicated. An individual scenario will typically: a) involve multiple device components; and b) bear on multiple potential issues. With careful design these may be configured to test key issues of interest, where multiple aspects of performance are providing important data.

Short simple scenarios may be appropriate for some evaluation goals while others may benefit from highly naturalistic, complete flights. Both types can provide tremendously valuable information about dynamic use and are valuable in combination. Considering alternative simulation approaches initially, such as integrated long scenarios versus separate short snippets, can help design the most informative evaluation for the least expenditure of effort and money.

Use of representative pilots is important. Successfully supporting test or other high-skill pilots should not be assumed to demonstrate that the device will provide adequate support with the target population. For a particularly critical issue that needs to be evaluated particularly accurately, it may be more efficient to have a large population of pilots doing a brief specific task, than to embed the critical task in the context of a long, complex simulation scenario, which can only be run by a relatively small number of pilots.

## 9. Evaluation Methods for Flight Test Phase

The goal of flight test should be to determine that the built device functions the same as the designed device, which has already been vetted. Flight testing is a very bad phase for discovering or assessing human factors problems with the interface. As with identification of any problem in flight test, discovery at this point has very high costs, relative to earlier discovery. More specifically, resolving issues prior to flight test is particularly valuable for assessing issues concerning support for attention, awareness, and understanding because these are very hard to see in a flight test context. When performance fails in flight test, it can be quite difficult to identify the basis of the problem, and whether issues of attention, awareness, and understanding are implicated. However, if performance is successful across the appropriate range of situations, this implies that these processes are appropriately supported. Thus, flight testing should provide reproduction of successful performance measured in simulation or earlier evaluations.

Flight testing relative to simulator testing is more limited in scope: dangerous scenarios are limited. Its scenarios are less controllable: weather or air traffic cannot be reproduced in flight test. Scenarios used in earlier testing of the device, typically in the simulator, can be adapted for use in flight test.

Scenarios should focus: a) where there is any concern about departure of the as-built system from the earlier, tested designs; or b) on key situations where there is doubt about the simulator's realism. For example, flight testing provides interaction with the actual National Airspace System (NAS) infrastructure. An additional limitation of flight testing is that the pilot is necessarily an FAA test pilot to minimize risk in carrying out the required maneuvers. Thus, the pilot in a flight test is far from a representative line pilot.

Flight testing relevant to human factors should be supported by an observer. Particularly for issues of attention, awareness, and understanding, one's own intuition and experience with a device is a poor measure of how well the device supports its intended functions; a pilot's report cannot include events they missed because of inadequate displays. The observer should be supported by a test script and scoring cards that identify the specific event types and pilot responses to record. Recording the data in flight is the priority to minimize information loss, and assessment relative to a pass/fail criterion can be determined post-flight as needed.

# 10. Summary and Open Issues

Failures of attention, awareness, and understanding have led to accidents and incidents, while successes underlie routine performance as well as "save the day" outcomes. Adequate support for attention, awareness, and understanding is a critical requirement for flight deck interfaces. This report provides information on methods to evaluate the adequacy of support for attention, awareness, and understanding in the context of certification by the FAA. Prior sections of this report provide an "encyclopedia" of methods for evaluating how well attention, awareness, and understanding are supported. Here we list several strategic principles, comment about tactics for useful evaluation, and mention some limitations in the scope of this report.

## 10.1. Strategy for Evaluating Devices for Adequate Support of Attention, Awareness, and Understanding

### *10.1.1. Start Evaluation Early in the Development Process*

The earlier a problem or potential problem can be identified, the easier and lower cost its resolution can be. Early involvement of the FAA is highly advantageous as this allows sharing of FAA human factors expertise with an Applicant and enables early identification and remediation of possible certification and safety issues. Inspection-based methods from design reviews can provide guidance before any prototype or simulation exists and before performance-based methods are possible. Certainly, early inspection cannot identify, let alone resolve, all issues. Yet, inspection-based methods can provide excellent value because they can be collected so early in design.

### *10.1.2. Rely on Performance-based Methods Early and Extensively*

Performance-based data comes from people carrying out tasks, not just from an inspection. This requires that the device being certified exists in some form that can be used in a task. This can be a device prototype, the device as part of a high or low fidelity simulator, or the actual built product. Prototypes and simulators can be used early in development.

- Use prototypes and simulation extensively, starting early in development. A great deal of useful performance data can be gathered from tasks with a device prototype or a simple simulation. Data from tasks that are just part of flying an airplane can still be very useful in assessing issues, such as having pilots use a device prototype or model to carry out specific functions it is intended to support. Data from prototypes may resolve design concerns

early. Resolving complex issues such as interaction with other flight deck displays may require later data collection from tasks carried out in a more extensive simulator. Bench testing can be done before integration into the flight deck. Different levels and types of test-bed fidelity are relevant for different issues and for different devices.

- Use flight testing to assess whether the system is built as designed, not to assess design issues. Flight testing has critical limitations:
  - If testing here does reveal design problems in hardware or even software, the cost of change is very high. Even if a change is required, the result is likely to be both more expensive and less safe than if addressed early.
  - Many conditions cannot be evaluated in flight test because they are too risky or because the conditions cannot be generated at will (e.g. wind shear).
  - Flight testing is necessarily carried out by test pilots who are not representative of the pilots who will be flying the airplane. Flight testing provides a vital final safeguard and may reveal unsuspected problems. However, work to resolve all attention, awareness, and understanding issues prior to flight test. Not only can simulator testing be done earlier in development, it is also less expensive, more flexible, and less dangerous than flight tests.

### *10.1.3. Rely Heavily on Typical Customer Pilots in Performance-based Evaluation*

- Use typical line pilots to understand how the device being certified will support operational attention, awareness, and understanding.
- It is increasingly recognized that test pilots are a very different population than line pilots and the difference is particularly critical for attention, awareness, and understanding. Test pilots may be able to generalize from their own experience to the experience of less skilled pilots on some topics, say, handling difficulty in manual flight; such projection, however, is not reliable for many issues in attention, awareness, or understanding. People have very poor ability to assess what they are or are not aware of, and even less to project from their own experience how a device might support attention, awareness, and understanding of a different type of person. Direct data from line pilots is critical.
- Do not substitute “in house” pilots working with the Applicant for line pilots in testing. Pilots who are particularly familiar with the device and its development are not representative of line pilots. While useful information can be gained from a variety of pilot types, testing the device with typical line pilots is critical.
- The importance of relying on performance data from line pilots is another reason for reliance on simulator testing: Here, but not in-flight testing, line pilots can be used.

### *10.1.4. Plan to Target Attention, Awareness, and Understanding Issues that have Greatest Threat to Safety*

Direct concern with inadequate support of awareness is relatively new within the certification process. Easy-to-apply rules, as are found for issues concerning font or glare, are not available. An effective, efficient evaluation provides information that reduces risk at a practical cost. Doing this for the more cognitive topics as in attention, awareness, and understanding requires identifying plausible threats and organizing the evaluation to target those. Tactics to help in this process are summarized next.

## 10.2. Tactics for Evaluating Devices for Adequate Support of Attention, Awareness, and Understanding

This report provides resources to help with design and execution of evaluation plans addressing support for pilot attention, awareness, and understanding. It provides a topics list for identifying specific attention, awareness, and understanding issues for a specific device; the issues list is treated more broadly in Report 5. It provides an organizational framework based on phase of development to aid development of maximally efficient evaluations.

### *10.2.1. Use an Issue-driven Approach to Design the Device Evaluation*

An efficient, effective evaluation depends on tailoring to the specific case. The principles and requirements in Installed Systems and Equipment for Use by the Flight Crew, AC 25.1302-1and rule CFR 25.1302, for example, are stated at a somewhat abstract level and require thinking about how they should be applied in a specific case. By identifying the issues or topics that pose the greatest potential risk, the evaluation can efficiently focus resources where they can do the most to reduce risk and lead to successful certification.

- Identify possible issues by referring to a list of human factors issues. Using a list such as that outlined in Appendix E and detailed in Report 5 flags high-level human factors issues, including those related to attention, awareness, and understanding. This identifies problems or inadequacies to look for in the device. Using a prompt like this helps identify specific instances of the general issues on the reference list in the device being certified, which form the set of active issues of concern.
- Track and update the active issue set. An initial set of specific issues can often be identified in an early design review and used by the Applicant and FAA to plan the evaluation. Issues can be dropped from the list when further review and analysis or additional data show the specific item is not a problem or when a design modification addresses the problem. Also, issues may be added that were not visible from inspection, as the system is developed and performance data collected.
- Target the evaluation on the issues. Because there are a very large number of possible issues, an efficient evaluation needs to focus on the issues that may pose substantial risk.

### 10.2.2. Pick Specific Scenarios, Situations, and Events that together Assess the Issue

Use scenarios that can diagnose an issue or set of issues. Do not focus just on generic normal and ‘worst case’ scenarios. Find situations that are most likely to show whether a specific aspect of supporting attention, awareness, and understanding is or is not adequate.

### *10.2.3. Ensure that the Methods are Relevant to the Issue of Concern*

Choice of relevant measures can greatly increase the efficiency and effectiveness of evaluating concerns, reducing risk, and moving the certification process forward. Many useful measures and ways of collecting data are available. Measures and methods traditionally used in evaluation can be extended by methods specifically developed to measure attention, awareness, and understanding. These specifically targeted methods can be adapted to the evaluation of devices and provide specific information about possible threats from inadequate support.

A poorly designed or “general purpose” evaluation may end up being expensive but do little to provide the information that will reduce risk or move forward in certification. For example, if the

issue concerns awareness about some system state or problem, make sure awareness is assessed, e.g., use verbal report; correlational measures such as eye-tracking or retrospective report may be helpful if general patterns of attention are of interest, but they do not provide definitive information about momentary awareness.

#### *10.2.4. Use Multiple Measures When Feasible*

Multiple measures are often very useful without adding much cost. For example, suppose that mode understanding is an issue and simulation scenarios are designed to test specific gaps in awareness of mode behavior; it can be very helpful to measure both the pilot's control actions in the moment and later verbal reports about their intent and their knowledge about mode behavior.

#### *10.2.5. Apply Human Factors Knowledge and Resources to Shape the Specific Attention, Awareness, and Understanding Evaluation*

There is no one test recipe to ensure adequate support for attention, awareness, or understanding. Rather, a good evaluation depends on identifying the most important issues and selecting methods appropriate to the issues. This report makes suggestions for how to identify issues and how to select the relevant and feasible scenarios, measures, pilot participants, and test environments. Because there is no "one size fits all" evaluation, available staff with human factors expertise are particularly valuable on both the FAA and the Applicate side of the certification process. Expertise is required to assess both the level of scrutiny required and how to provide this effectively and efficiently.

#### *10.2.6. Anticipate Increasing Need for Evaluation of Support for Cognitive Aspects of Performance*

The complexity of airliner systems will continue to increase. Support on the flight deck to ensure that pilots can effectively understand and manage those system will become increasingly important. Post-design support from training, procedures, or briefings may become limited in its ability to provide effective remediation for marginal design choices.

### 10.3. Scope Limitations

Many important issues fall outside the scope of this and our related reports. Very broadly, we recognize that considerable practical experience is important for putting to work the information provided in this report. This is true for certification for human factors broadly, and particularly for attention, awareness, and understanding issues. We discuss two broad types of limitations: managing the complexity of evaluations and managing relations between Applicant and FAA parties.

Evaluation in the context of certification is a complex process, undertaken with limited resources. There is no fixed recipe. Rather, evaluation in the context of certification requires applying general methods as appropriate to the specific case.

- This report provides information about best methods for evaluating whether a device undergoing certification provides adequate support for issues concerning attention, awareness, and understanding. With strategic planning, best-practice methods can also be efficient. We recognize that there nevertheless are tradeoffs in what would be desirable and what is feasible given resources. Informed choices depend on understanding the advantages of various general methods and the importance of the various issues relevant to a specific case.

- This report is organized by development phase and by issue. We introduce a general method and then our examples illustrate testing a single issue at a single phase. We think this is the clearest way of introducing the methods. Thus, the report describes the "building blocks" for constructing the overall testing of a device for attention, awareness, and understanding issues. The space of combinations of multiple methods testing multiple issues together is very large. We do provide some comments about how testing of multiple issues might be integrated or how later phases can benefit from the results of earlier testing. Nevertheless, how these "building blocks" are effectively and efficiently integrated for an overall certification evaluation of attention, awareness, and understanding issues must be addressed and refined case by case.
- We recommend that the evaluation be focused on reduction in safety risk, and that organization by issue is an effective method to do so. Certification grounds out in the particular Part 25 rules under which the device is being certified. This report does not discuss how to map issues to rules, but suggestions are made in Mumaw, Haworth, Billman, & Feary (2019).
- This report addresses evaluation of attention, awareness, and understanding issues in isolation from evaluation of other issues, even other issues within human factors. We believe that there are multiple opportunities when one testing sequence can assess multiple issues. Currently, one simulation-based evaluation often does test multiple issues. Exploration of when and how best to design one test to diagnose multiple issues is an important open question.
- Setting criteria on performance will depend on what is operationally relevant for the specific device and task. Frequently, safety critical actions depend on correct awareness of values and on understanding of key aspects of the situation. In these cases, very high criteria will generally be needed. Criteria are linked to sample size. If one of four crews experience a problem, this might be interpreted as an outlier or as suggesting a quarter of crews will have problems. How to provide guidance for setting criteria that is both useful and generally applicable across devices and issues remains an important open topic.
- Aviation systems are increasingly complex and increasingly depend on interaction among and integration of components. Not only are systems integrated, but the integration is typically built up and revised over the course of development. A complex problem, such as certification of a complex system, needs to be decomposed some way to make it tractable. For example, we recommend decomposing the problem by phase and by issue. However, ensuring that the problem is solved when the pieces are put back together is difficult. New approaches may be very valuable.

Designing, conducting, tracking, supervising, and assessing the evaluations in a certification project requires significant interactions between the FAA and the Applicant. Both the context of the evaluation and the roles played by FAA versus Applicant personnel is open to some negotiation.

- An Applicant has, we believe, a strong self-interest in carrying out relevant, early testing to minimize requirements for costly, late change. However, management of testing cost is an important pragmatic question. An important open question for the FAA is how the FAA can encourage, specify, or reward relevant, early evaluation by the Applicant. One question is whether and how "credit" might be given for success in earlier phases of testing that can reduce the burden of later-phase, higher-cost assessment. We do not address how to manage pragmatic tradeoffs or what might incentivize or inform testing early in the development process.

- We focus heavily on emphasizing the importance of early assessment. Currently, the Applicant executes early evaluation, and FAA executes the final flight-test phase. FAA involvement in early assessment is likely to be highly beneficial and developing approaches for early involvement continues to be an important topic.
- The on-going process of relationship-building between FAA and Applicant personnel is clearly important. Relationship-building aims to develop buy-in to early-phase assessment broadly, which is particularly needed for assessment of support for awareness and understanding. Hopefully, the methods outlined in this report will assist the FAA in that process as well as more directly suggest the types of evidence that are sufficient to address many issues concerning awareness and understanding.
- FAA resources (as well as the Applicants) are limited. Specifically, the FAA has limited human factors staff available to support the certification process. This limited staffing is an important factor necessarily impacting the process.

# Appendix A. Illustrations of Issues Evaluated in Design Review, Prototype, and Simulator-based Phases

The body of this report is organized by types of evaluation methods useful in each of the possible evaluation phases for issues concerning attention, awareness, and understanding. Appendix A provides a complementary perspective by sketching how methods can be applied to issues concerning four broad requirements the flight deck should meet. Across the intended functions, the flight deck as a whole and its component devices should:

- provide information the pilot is looking for
- ensure provided information is accessible and manageable
- direct pilot attention to information about important changes
- support situation understanding and assessment of action in the operational context

These broad functions provide a framework for identifying specific issues concerning attention, awareness, and understanding.

Identifying the specific potential issues and vulnerabilities of the device undergoing assessment is a very important aspect of the certification process, as identifying these issues ensures that evaluation resources will be spent where they can best reduce risk. Mumaw, Haworth, Billman, & Feary (2019) provides an issue-centered approach and gives a accounting of issues that may be of concern. Mumaw, Billman, & Feary (2019) provides guidance about scenario design.

We include more examples from the earlier phases because assessment here may be less familiar and less used than in later phases. Further, it is extremely valuable to identify and resolve issues as early in the process as possible.

## A.1. Design Review: Example Evaluations Organized by Issue

### *A.1.1. Provide Information the Pilot is Looking For*

*Example DR1: Needed information is missing.* Lack of indicators that provide needed information can be a serious design flaw and design review provides an early opportunity to identify this. Needed information may "fall through the gap" if the design assumed relevant information would be provided appropriately in other flight deck components but this is not the case in the current context; information may not have been recognized as important or relevant because of changing needs or because of inadequate analysis of the functions relevant to the device. Someone with human factors expertise may be able to identify what information is needed but is not provided by any indicator. Identifying "edge case" usage and attempting a cognitive walk-through for such cases may reveal gaps in the information coverage. (Redundant information, rather than missing information, is usually not such a serious problem for safety.)

*Example DR2: Perceptual representation of the information is inadequate.* Rules and ACs provide considerable guidance on many perceptual aspects of displays that impact the ability to encode it, whether by seeing or hearing. Review of design sketches or specifications may allow early discovery of basic problems with the how information is represented, particularly at the level of how individual indicators present information:

- Size, font, resolution, and color may all affect readability of text or symbols.
- Volume and clarity all affect audibility of aurally presented information.

Problems with how individual variables are represented, and where design of indicators conflicts with the many rules specified at this level, may be identifiable by inspection. “Fresh eyes” that were not involved in creating the design may be able to spot problems by reviewing the design specifications. While more difficult to do based on partial design specifications, consideration of how displays of multiple indictors may interact may be useful. For example, a less important variable may be displayed in a more salient manner than a more important variable. Information needed together may not be available together.

*Example DR3: Labels and icons may be hard to identify or understand.* The selected words and icons can be inspected to assess whether they are familiar, consistently used across the interface, and in accord with standard practice within the industry. Color use should be similarly consistent internally and with industry standards as well as with FAA rules.

*Example DR4: Context variables may be missing or inadequately displayed.* This may be more difficult to assess from a design review but it may be very valuable to identify any problems as early as possible. Indicated current values of variables may be confusing or hard to interpret if displayed without context, when display of context information itself is confusing, or when the context change is not well represented. Context includes static and dynamic aspects, such as the current general situation, current weight, items on the MEL, and optimal cruise altitude. The context itself may be characterized with variables such as current thresholds or current expected values.

Important context for many indicated variables includes the following:

- Its expected or normal value (when it can be determined)
- Its commanded value (when there is one)
- Its desired value (such as clearances)
- Any relevant thresholds (e.g., the upper bound of normal range, operational decision points) that may be near the current value or that represent an abnormal state

Information about historical or predicted change can be very valuable:

- Its change over a short period of time, such as increase, decrease, and rate
- Its history of change over a longer period, (e.g. whether fuel is dropping as expected over an extended period)
- Its estimated time to reach a commanded or expected value

These context variables typically are important variables in their own right, they need to be appropriately displayed at an individual level, and the overall display needs to be organized so comparisons between current values and context values are easy to make. It may be useful to directly display relations so the relations can be seen rather than computed (e.g., distance from stall speed, distance from waypoint).

### *A.1.2. Ensure Provided Information is Accessible and Manageable*

*Example DR5: Display navigation paths may be unclear, too long, or poorly organized.* Due to the limited amount of display space and the large amount of information to display, flight deck components are often designed to reallocate the same display space by navigating to different information, often through menus. The clarity of organization and of symbols used in display navigation can be inspected. The design of display management symbols may be inspected for

understandability and consistency as recommended for indicators. The menu organization should be assessed for how easy it is to move from viewing one set of indicators to viewing another. The length of navigation may be inspected, such as counting the number of clicks or other operations needed to move from accessing one variable to accessing another that might be desired in the same context. Navigation paths should be traced among variables that are commonly used in the same situation and traced among variables that might be used together rarely, but in critical situations. Even in design review, thinking in terms of the scenarios of use can be very helpful. "Edge cases" and non-normal situations may be particularly useful to inspect. For example, these might include situations requiring unusual combinations of information, hence unusual navigation paths, or reviewing the effects of total failure of each display component. Ideally, information that is used together can be displayed together. Further, no information should be "buried below the surface," in the sense that unclear or extensive actions are needed to view the information. Problems in accessibility may be particularly likely when new displays are added to an existing set of displays. The new display may not fit naturally into the existing menu hierarchy, and navigation to the display may be less intuitive.

*Example DR6: Display management methods may be inadequate.* Navigation though a menu structure is a common form of display management. Additional methods for configuring displays can be helpful, particularly if automation may also change how or what information is displayed. Methods for reconfiguring displays, to group information or to prevent change from a particular configuration, may be helpful. Display strategies might be considered, such as whether pop-ups should be used at all, given that they occlude information. The controls provided for making changes can be reviewed. Reconfigurable displays can provide a useful "scratch pad" e.g. supporting comparisons in decision-making. Flexibility as well as transparency of information access should be considered.

*Example DR7: Data validity is not indicated.* If uncertainty about data validity is not presented to the pilot when available, this lack reduces pilot ability to reason from the data. Some indications are derived from multiple sources, by a voting or other integration scheme. If one of the data sources is suspect, it may be valuable to communicate this information to the pilot. The design can be inspected for where and how uncertainty is represented.

*Example DR8: Transient information is not adequately preserved.* Most indicators present information about the current state, which is dynamic. However, some information must be preserved long enough for both pilots to process the information. This may be an issue for visually presented information and is one of the reasons it can be valuable to preserve information about previous states. The auditory modality is inherently transient, but is used for very important information, as in alerts and non-normals. Further, crew often want to clear auditory signals quickly. A design review can identify whether or not methods for preserving transient, and particularly auditory, information are provided.

### *A.1.3. Direct Pilot Attention to Information about Important Changes*

The primary method of directing pilot attention is through alerts or other manipulations of salience. Other indicators of change may be included, such as a green box surrounding a recently changed autoflight mode. Because effectively directing attention involves tradeoffs among many factors, it may be hard to tell much about alert adequacy from inspection of the design. Design review may be useful on some issues, as well as identifying what elements depart most from previous, acceptable

designs. However, inspection may say very little about other issues, such as whether one alert is likely to mask another.

*Example D9: No alerting scheme was used for an important change.* A design inspection can give an accounting of what events are intended to trigger which alerts. This accounting can be reviewed to identify any events that should gain the pilot's attention but lack a method for specifically drawing attention.

*Exampl eD10: The alert was not understood.* Alerts might be reviewed for use of standard terms and for how clearly they refer to the particular alerted event. Further, the alert-as-written can be reviewed for what information it provides about whether some action needs to be taken.

*Example D11: The interface does not prioritize multiple alerts.* Temporal order of occurrence is a poor default form of prioritization and the design may be inspected for whether an alternative prioritization scheme is provided. Some forms of cognitive walk-through may be helpful to identify system behavior when multiple alerts occur. This style of evaluation will be most beneficial if the process of inspection considers scenarios that elicit multiple alerts, and particularly, unexpected combinations of alerting events, so that an inspection reveals how these are handled.

### *A.1.4. Support Situation Understanding and Assessment of Action in the Operational Context*

It may be very difficult to evaluate how effectively the interface supports the pilot's situation understanding and assessment from a design review. Nevertheless, concerns for these topics may emerge from review of other issues. Of course, problems with information presentation are likely to lead to problems in how the information is used. If the device has a bounded, specific function of supporting a higher-level decision-making activity, it may be possible to do a useful cognitive walk-through of how the decision aid would work in a limited range of situations. For example, for a decision aid for selection of alternative airports, it might be useful to review the design for how it would support this type of decision. However, most assessment of these issues will likely depend on collecting performance data, either with prototypes or in a simulator.

## *A.2. Prototype Phase: Example Evaluations Organized by Issue*

Awareness of relevant information frequently is not viewed as a task on its own but as contributing to multiple tasks. Nevertheless, it can be very effective to assess awareness in relative isolation, typical in prototype-based assessment.

The purpose of the examples here is to show how the general method can be applied to a variety of cases, spanning the four broad topics concerning attention, awareness, and understanding. The method implemented will of course be tailored to the specific case; it will depend on what aspect of the interface is the focus of investigation, and what conditions will best identify potential problems.

The examples are described as assessing how well a particular interface supports the participant in some task measuring or depending on awareness. In addition, each of the examples and methods can be very usefully applied to compare two alternative designs. This can be particularly useful as formative evaluation, helping an Applicant identify a more effective design choice. It can also be used to compare a proposed design to an established design, if comparable functionality exists. Such comparisons can be particularly valuable where sharp performance criteria are lacking, as will typically be the case for prototype-based assessment.

### *A.2.1. Provide Information the Pilot is Looking For*

*Example P1: Assessing whether one indicator for a specific variable may be unclear or problematic.* If a new display concept is being developed with novel icons or labels, early performance testing may be very helpful. While the design review may eliminate certain problems, it can be useful to test actual clarity with performance data. If only one element is new or has suspected issues, it may be efficient to test only that element. The goal is assessing whether participants can accurately read and report the displayed values without confusion among the values displayed for this variable and without confusion with values of other variables. To conduct a useful assessment, the prototype needs to be developed to the point of producing static images showing a wide range of the displayable values. A large number of displays can be produced relatively quickly from such a prototype.

Typically, a small number of participants, perhaps 6–12 can be used, each contributing data on a large number of quick trials. Participants should be familiarized with what the display can show, and what the appropriate response is for different values. For example, if icons or abbreviations are used, an easy, meaningful, standardized label should be used. If continuous variables (particularly with analog representations) are used, degree of precision in reporting should be defined and familiarized. If there are a small number of possible variables, attention should be paid to minimizing predictability of values and controlling effects of guessing. It is useful if presentation of a series of images can be controlled automatically, if only though a slide-show presentation. If there is only one variable presented, the participant always knows what variable to report. Several presentation-response modes are possible. Each image can be presented for a fixed, brief exposure time, and the participant asked to report the value as soon as possible after the image is presented. Alternatively, the participant can be told to look for a specific value occurring anywhere within a series of rapidly presented images, and to report whether or not the target occurred anywhere in the sequence. This can be particularly helpful if there is concern about confusability in how certain values are represented.

*Example P2: Assessing whether a set of indicators for multiple variables may be unclear or problematic.* If there is interest in multiple indicators, it may be efficient to test displays of multiple variables together. In the simplest extension of Example P1, presentation of each variable can be blocked, and at the beginning of each block the participant told the target variable to report. Across displays the values of the nontarget variables can change, creating a somewhat more complex context. This design may require a somewhat longer orientation phase and pausing between blocks, but otherwise follows the logic of Example P1. If multiple variables are of interest this design is efficient because participants need be scheduled only once. In addition, the presence of multiple, changing displays provides a richer and more realistic context. Patterns of errors may be particularly informative in identifying possibly confusable states. Even if the focus of investigation concerns correctness, it can be informative to collect response times, as speed is typically a more sensitive indicator of problems than is error rate.

In a simple extension of this design, the target variable changes from trial to trial, rather than remaining fixed within a block. Here the participant is told in advance of each trial what to look for. This procedure includes effects of attention switching as well as looking at the relevant variable, identifying the value, and reporting.

*Example P3: Assessing whether design elements within the new device conflict with existing flight deck components.* In some cases, there may be issues about confusability, redundancy, or interference of the new device with specific, established flight deck components or devices. Prior to evaluation with operational tasks and an advanced simulator, it may be possible to include selective components as context. This is likely to be feasible for static displays. The relevant components in the old flight deck can be included in the prototype mockup with the components of the new device, with changing values inserted in the display. Here variables shown in the old components may be probed as well as those in the new. Static displays will not address questions of interference among components as deeply as can be done with a dynamic simulator, but this very light weight method may be very useful, for example, in providing early performance data assessing concerns about confusability or interference between new and old elements.

### *A.2.2. Ensure Provided Information is Accessible and Mmanageable*

*Example P4: Is display navigation and finding relevant variables from different starting states clear and quick?* Displays are often quite complex and not all variables are visible at once. Rather, the pilot must configure displays to show the desired information by operations such as navigating through menus, selecting windows, scrolling, or rescaling. There may be particular transitions or navigation paths that are suspected of being problems, or a broader sampling of navigation paths provided (including detection of unintended paths) may be desired. Ease of accessing needed information can also be evaluated with quite early prototypes, such as mockups in display software. Once it is possible to click, scroll, or carry out the display management operations, it is feasible to assess access. Early assessment may be based on performance such as number of operations to access the target information from differently configured starting points.

*Example P5: Display management methods may be inadequate.* As with many issues where design review is helpful, it can be easy to miss problems that emerge in interaction. Navigation paths can be evaluated with a prototype that allows stepping through specified transitions. A large set of paths can be evaluated by specifying an initial configuration and the information to locate. Backtracking or a large number of clicks can indicate problems, in menu labels and/or menu structure. More common paths may be designed to be shorter. A large amount of guidance about menu design and evaluation is available from general HCI research. Flexibility as well as transparency of access should be considered

### *A.2.3. Directs Pilot Attention to Information about Important Changes*

The effectiveness of support for noticing changes is strongly affected by context and the specific dynamics of change. Assessment depends on the availability of some form of dynamic prototyping and is particularly valuable to reassess in a simulator. Nevertheless, data-driven assessment can be an important improvement beyond design review. Using secondary tasks to add workload may be helpful in these types of assessment.

*Example P6: No alerting scheme was used for an important change.* Does a particular change need an alert or a more effective way of helping the pilot notice a change? This type of focused question may be investigated if a dynamic prototype is available. If a design goal is that a change be noticed within a certain time, some assessment can be done with a prototype. For example, the participant may be asked to report when any of several values change, and time to report the critical variable can be evaluated. Comparison between different methods of signaling change can be made this way, for example, a comparison can be made among options such as no extra indication that the flight

mode changed, a green box, or a flashing value; alternatively, different durations of the change-marker may be compared.

*Example P7: Is a particular alert sufficiently salient to draw attention and allow understanding?* Once dynamic prototypes are available some assessment of salience of alerting, or other methods of drawing the participant's attention can be made. Complex displays are desirable for doing this. Following an alert, the display can be covered or blanked, and the participant asked to report the alerted value or information. Slow or inaccurate reports can identify problems. Alternative alerting schemes can be compared.

*Example P8: The alert was not understood.* Alerting assessment is better done in the context of some larger reporting task, so the participant's attention is not exclusively focused on the alert and some element of surprise is involved. When an alert occurs, the participant is asked to report what the alert indicates. This may include information about higher level status or aircraft behavior, or what actions should be taken, as well as indicating a specific value is out of range. These issues likely require pilots as participants, since understanding the alerts may draw on knowledge of piloting.

*Example P9: The interface does not prioritize multiple alerts*. While inspection can reveal some problems with prioritization, such as lack of any prioritization method, it is also valuable to collect performance data once prototypes are available. This can target areas where there are concerns about the design or provide a broader assessment of the prioritization scheme. Scenario design is particularly critical as alert prioritization is most important in situations of multiple and potentially interacting faults. Tasks can be designed where scenarios are presented, and the user is asked to identify what is the most important issue or the most important next actions. Next-actions might include the option of seeking additional information to clarify the current situation. This can be done prior to development of a prototype or simulator that allows the pilot to actually issue control commands, respond to the alerts, or reconfigure displays.

### *A.2.4. Support Situation Understanding and Assessment of Action in the Operational Context*

The prototyping phase allows some exploration of how effectively the interface supports the pilot in building an integrated situation model, predicting future states, and assessing possible actions. Pilots are the most appropriate participants in evaluating these issues. Useful information can often be gained even from static prototypes by presenting a series of images showing a sequence of states and asking about the situation shown. However, if dynamic characteristics are an important part of the specific concern, dynamic prototypes and simulator-based test environments become more important. Specifically, if both a changing environment and time of pilot response are important evaluation concerns, the issue is better addressed in a dynamic test environment.

*Example P10: Interface does not support fluent updating of an integrated situation model.* The pilot needs a range of relevant variables to include in their situation model and maintain situation awareness; the key information needs will vary with situation. Fluent updating of the pilot's situation model may be hard to assess without a fairly advanced ability to work with dynamic scenarios or simulation. However, earlier, prototype-based assessments can be done to look for and capture any information gaps for the scenarios presented. Information gaps can be identified if pilots are seeking or capturing information from other sources. This might show up as a request for information, reference to external documents, or generating notes for oneself e.g., to preserve information about a

prior state or pending action. Lack of fluency may be detected not just by speed, but again by looking for reliance on external resources, to make comparisons or compute values.

### *A.2.5. Future Events Cannot be Projected*

There may be concern that the scope or presentation of the information provided is inadequate for the pilot to accurately project what will happen. For example, different modes of the autoflight system may produce different behavior that is hard to project from indicated information; or, the pilot may need to project changes in fuel burn due to changes in wind or in cleared altitude. An initial method of investigation can be to provide a sequence of images showing the relevant device indications over a period of some change, and to probe the participant about the future. Alternative display designs can be compared. Comparing performance with alternative designs allows assessing, for example, whether projection is much better if particular information is included in the displays. Participants can be probed by asking them to describe the current state including current processes, what will happen next without pilot actions, the basis or explanation for these answers, and whether there is other information needed to make good projections. Thus, it may be possible to collect valuable information about projecting future events without animation or dynamic prototypes, through careful design of scenarios and displaying a sequence of static displays.

*Example P11: It is difficult to generate or assess alternative possible actions or difficult to weigh cost/benefits of alternative actions.* Investigating how adequately an interface supports the pilot in assessing alternative actions can be done similarly to investigating how adequately the interface supports projecting future events (Ex P10). Indeed, both questions can be investigated together by including questions asking the pilot what if any pilot-actions are needed. Assessing how well the interface assists the pilot in weighing alternative actions in a particular scenario can be done multiple ways. The pilot can be asked to generate alternative actions appropriate in context and then to evaluate their relative risks and merits. Alternatively, the pilot may be presented with the alternative actions for the scenario and asked to evaluate them. While it is possible to evaluate multiple aspects given a particular scenario, care must also be taken not to ask too many questions of one person for one scenario lest the question-answering process becomes burdensome and unnatural.

## A.3. Simulator Based Phase: Example Evaluations Organized by Issue

Because simulator testing is relatively expensive, it is often useful to use scenarios designed to assess multiple issues or concerns for the device or set of devices being evaluated. The illustrations here are by issue, but multiple issues can be assessed at once, though this requires care in designing the scenarios. For simple issues, shorter or more repetitive scenarios may be adequate. Ensuring that the pilots have sufficient time to orient to the scenario can be valuable. However, “stress testing” where pilots are not optimally oriented (as they are in training simulation runs) can also be valuable. This may better reflect realistic flight settings in which a pilot’s attention may not be optimally oriented and engaged at every moment.

### *A.3.1. Provides Information the Pilot is Looking For*

*Example S1: Context variables may be missing or inadequately displayed.* Is all the needed information available and can it be easily accessed in operational situations? It can be useful to design a prototype and simulator test as a pair. Basic availability of information can be checked even in a static prototype by showing displays of the flight deck in key states and asking some participants simply to read off requested values and other participants to carry out projections that use the values. Suitability in an operational context can be tested in a (dynamic) simulator. The evaluation context

should ensure that for some cases the pilot is not expecting what and when particular context information will be needed. This is typically done by nesting the key scenario within a longer flight sequence. Participants should be representative line pilots. Fidelity of the simulator will depend on the specific issue but for many issues concerning awareness and understanding a mid-fidelity simulator may be sufficient. The visual context, however, will typically need to accurately represent both displays and out-the-window view.

This example case concerns use of context variables to anticipate autoflight behavior and feasibility of planned flight path. Management of the autoflight system depends on context variables such as tail wind and predictions for making crossing restrictions. Assessing whether the context variables are adequately displayed for effectively accomplishing goals can be assessed in a simulator. Diagnostic scenarios could include situations such as this: The cleared STAR approach has some steep descents between waypoints, with some altitude requirements provided as windows and some as a fixed altitude. There could be situations where flying the bottom of the window on an earlier waypoint would enable making the second but flying the top of the window would not. The flight has been cleared for approach on the STAR, there is a merge waypoint ahead, and traffic on both own and merging paths is heavy. In this situation the pilot will need to integrate information about autoflight mode, waypoint altitude and speed restrictions, tail wind, and likely traffic separation to assess whether the airplane will be able to fly the STAR. Shortly before being probed, the tail wind component of windspeed could increase to a problematic level meaning that the airplane probably cannot “go down and slow down” to meet one of the upcoming waypoint restrictions. The timing before the probe is set to an operationally meaningful increment, such that the pilot should check windspeed within the increment. A performance criterion for each component of the task should be set in advance if at all feasible. At this point in the scenario, the pilot will need to recognize that the situation needs evaluation and decide what to do: defer decision; no action needed; fly manually and comply with the flight plan; or report unable and request an alternative. At the time of the probe, the screens are covered (or blanked) and the pilot is asked to report multiple relevant variables sampling from or including windspeed, current airspeed, current altitude, autoflight mode, upcoming waypoints with altitude restrictions. The critical context variable (e.g. windspeed) will be probed first, to minimize effect of interference from verbal report. If variables have been understood and integrated into a meaningful situation model, values will be reportable over a longer period of time. Following the focused probing of variable values, the pilot will be asked a series of more specific questions, beginning with open questions such as ‘how is the approach progressing?’, ‘will you need to use the speed brake?’, and down to ‘do you anticipate any difficulties meeting the waypoint restrictions on WayPt?’.

This scenario focuses on accuracy but also captures a threshold measure of timing. That is, it will assess whether the pilot was able to gather information within an operationally relevant interval. The scenario assesses both awareness of a key variable and use of that in decision making.

Evaluation Issues: Performance, specifically timing, will differ depending on whether the pilot is able to predict what variables will be queried. For example, asking the pilot about future flight path will likely increase pilot attention to and monitoring of this information subsequently. As in analogous cases mentioned earlier, an initial, unexpected probe may give “normal” speed and accuracy, while later probes may be looking at “best” performance.

Designing simulator scenarios to test multiple issues is valuable because simulation testing is expensive and complicated. An individual scenario will typically: a) involve multiple device

components; and b) bear on multiple potential issues. With careful design these can be configured to test key issues of interest, where many aspects of performance are providing important data.

### *A.3.2. Ensure Provided Information is Accessible and Manageable*

*Example S2: Display management methods may be inadequate.* As with many issues where design review is helpful, problems may be missed until a dynamic test-environment is available. Management of the display should ensure information in use is not unintentionally removed. For example, pop-up displays can occlude information under them, and may be a risky design choice. It may be very difficult to ensure that the occluded information is never needed with the pop-up until the device is in use in a dynamic environment (whether prototype or simulator).

### *A.3.3. Directs Pilot Attention to Information about Important Changes*

Simulator-phase testing may be particularly valuable for assessing dynamic aspects where time to notice and understand is particularly important. Assessing the scope of individual alerts in the context of the overall alerting may be very valuable.

*Example S3: The interface does not effectively prioritize and organize multiple alerts*. A comprehensive plan is needed to prioritize and manage multiple simultaneous or overlapping alerts. The interface needs to aid the flight crew in making sense of the full set of alerts. Highly salient alerts may prevent attending to other alerts that are occurring simultaneously or in close succession. In turn, this may prevent the pilot from being aware of and understanding the full set of alerted events. Evaluation may focus on both noticing and interpreting alerts. Scenarios should look for atypical but critical situations where multiple alerts will overlap or occur in close temporal proximity. Assessment might focus on an individual alert, but it may be most valuable to assess how the pilots are able in context to interpret and assess information from multiple sources including alerting. Both verbal report and IDAs may be useful measures. Effectiveness of an individual alert or a set of alerts can be assessed by the accuracy of understanding the meaning and implications for future capabilities of the airplane. Useful measures include: a) talk between pilots; b) control actions taken; and c) information-seeking behavior such as reconfiguring displays. Probe questions can be introduced as well; naturalism can be increased if the questions can come from ATC or from a confederate pilot.

Because alerts frequently occur in surprising or startling situations, repeating test events with the same crew is not a good idea. Responding will not be the same when the alert is expected as when surprising. For evaluation efficiency, assessment of alerts may be more efficiently done as part of scenarios assessing multiple evaluation questions.

### *A.3.4. Supports Situation Understanding and Assessment of Action*

*Example S4: Future events cannot be projected.* All displays can influence the pilot's ability to project future events. Some devices may particularly focus on supporting prediction, planning, and adaptation, and thus may raise issues about how effectively this is in fact supported. Any device or display has some core scope of intended use, whether predicting range based on fuel and cruise altitude or arrival time based on wind, weather conditions, and autoflight cost index. Assessing support for projecting the future requires creating situations where the relevant aspects of the future are: a) not predictable based on general expectations; and b) depend on information provided by the device. Depending on intended use, it may be informative to have the pilot do multiple, expected predictive tasks, since normal use of the device may not typically involve surprise or response to

alerts. This can assess the coverage, accuracy, and clarity of the support provided. The most valuable testing sequences may be sampling a wide range of conditions and looking at how well the pilots are able to predict future aircraft behavior across them. Pilots using the device may vary in their individual prediction skill; an interface that supports a task well will reduce if not eliminate these differences. Since prediction is likely a skill on which pilots vary, it may be particularly valuable to include a relatively large sample of pilots, who differ not only in general experience, but in the conditions they normally predict.

In addition to focused testing of intended use, it may be valuable to assess how performance on other tasks is affected. For example, does developing predictions using the device interfere inappropriately with other tasks likely needed in that context?

## Appendix B. Simulation Data from Human Performance Modeling Methods

Models of human performance and interaction produce data showing how the device is expected to work on some collection of tasks. A useful model provides a well-specified, possibly computerized, method for estimating how well a device will support some aspect(s) of its intended function. Useful models can be developed to aid assessment when one or more factors affecting effective use are well-understood so their impact can be modeled and predicted. To our knowledge, no model takes into account all factors known to influence a device's effectiveness; rather, models provide a systematic way of looking at a selective set of one or more factors. Thus, models can provide a limited but systematic component of assessment.

A model might characterize directly observable properties of the device (such as the minimum number of clicks to go from an initial display configuration to a configuration showing a particular target piece of information); broadly, this type of model simulates what an expert might do by inspection, and may be able to provide more extensive coverage than is feasible for an expert. Alternatively, a model might predict some aspect of performance data, such as the average time it will take a user to navigate from an initial configuration to the target information. A model might directly predict attention or workload, but include some assumptions about impact on observable behavior, such as more attention means higher percent of time spent fixating or more workload means higher chance of errors. Broadly, performance models simulate patterns expected if participants were measured using the device to carry out the modeled tasks.

Where assumptions of a model can be formalized, it may be possible to build and run a computer model that checks for these properties, rather than relying on expert opinion (for the first type) or on observed performance (for the second type). It is important to note that existing models only check for quite specific aspects.

Model-based assessment may be conducted at multiple stages of device development. Specifically, modeling candidate designs for specific factors might be possible, particularly for visual displays, before it would be possible to collect any user data.

There are existing modeling tools to predict the relative salience of display elements. With these modeling results, it may be possible to make a judgment whether the display elements that should most attract attention are the ones that are visually most salient. Intensity, color, and orientation dimensions—and contrast on these dimensions—are important predictors of display-based, or "bottom up," salience. Models for how these factors come together to produce salience have been developed in visual science (Itti, Koch, & Niebur, 1998) and implemented in MATLAB routines (Harel, Koch, & Perona, 2007). In one study, this type of model was used to predict eye fixations (Parkhurst & Niebur, 2005) which can be a precursor and predictor of attention. Some models specifically target attention (e.g., Wickens et al., 2008). Detailed review of modeling methods falls outside the scope of this report and modeling is not covered further.

## Appendix C. Eye Fixations and Other Eye Tracking Measures

Eye fixation, derived from eye-tracking methods, is another measure applied to assess awareness. This method records the user's eye fixations on the interface. Eye-tracking and the resulting sequence of fixations provide a relatively new method and new data type for use in interface evaluation, and have been suggested as a complementary measure of awareness (e.g., Wilson, 2000).

Indeed, eye-tracking offers several attractive properties:

1. It is relatively unobtrusive in terms of interfering with operational tasks.
2. It provides continuous data.
3. It does not require asking the pilot about their awareness of particular variables (which potentially might change the task).
4. It does not require identifying discriminative actions (IDAs).

However, it is critical to ask what we can infer about awareness from eye-fixations. The two are sometimes considered to be directly linked; specifically, it is sometimes assumed that any element in the world that is visually fixated enters awareness. However, this assumption has been shown to be incorrect. There is a large and growing body of evidence demonstrating that looking (fixation) does not always lead to seeing or awareness. Understanding the magnitude and scope of these attentional limits is critical for understanding the limitations (and strengths) of what eye-tracking can tell us.

Studies have shown that people can look at (fixate) a highly visible stimulus, even for several seconds, but fail to become aware of the information at the location, a phenomenon labeled *inattentional blindness* (Mack & Rock, 1998) or *change blindness* (Simons & Levin, 1997). While these terms refer to somewhat different attention limitations, their differences are not critical for our purposes. Specifically, in experiments using relatively simple dynamic or static computer displays, people are frequently unaware of an unexpected object in the field of view that is irrelevant to their task. For example, participants may be unaware of a red cross when counting the number of "bounces" of black or white circles and squares, or unaware of a "+" moving among L and T targets. This may occur even when the unexpected object passes through a fixation point and when participants are free to look where ever they choose (Beanland & Pammer, 2010). This phenomenon occurs in a wide array of settings. (See Most, Scholl, Clifford, & Simons, 2005, and Simons & Rensink, 2005, for a review and discussion focused on simple stimuli.)

These studies show that goal-relevance and user expectation are powerful factors affecting inattentional blindness. The likelihood of inattentional blindness to an object is much greater if the object is not relevant to the current task and if the object is unexpected. Inattentional blindness occurs more when a primary task is demanding, or the person's attentional resources are taxed. Exogenous stimulus properties modulate the magnitude of the effect, but mismatch with endogenous participant goals and expectations is a key driver.

Inattentional blindness has been produced in a wide variety of naturalistic and work-relevant contexts, as well as with simple stimuli. When watching video tapes of everyday events, people are frequently unaware of a variety of unexpected though highly visible appearances—e.g., a person in a gorilla suit walking through two small teams passing a basketball (Neisser & Becklen, 1975; Simons & Chabris, 1999). In live interaction, participants can also fail to see easily detected changes when the moment of change is briefly occluded. In one study, roughly half the participants failed to notice a switch in one of the people in a conversation when a one-second gap occluded the switch (Simons & Levin, 1998). Studies directly relevant to work settings can be done in the laboratory, using

micro-worlds to simulate a naturally occurring task. For example, Vachon, Vallières, Jones, & Tremblay (2012) used a micro-world task based on detection of hostile aircraft, where detection was measured by timely response to the threat. Detection was indeed much more likely if the aircraft had been fixated promptly after the change (78.1% detection given fixation) than if it had not been (7.2% given no fixation). However, fixation was no guarantee of detection, as 17.9% (100–78.1%) of the undetected changes nevertheless were fixated both before and after the change. In the domain of evaluating general-use human-computer interfaces, Varakin et al. (2004) report related cases of users "looking at without seeing" large changes on their screen. The extent of change blindness is large and counter-intuitive; indeed, participants asked to predict the prevalence of change blindness reliably and dramatically underestimated the occurrence of change blindness (Levin et al., 2000). Thus, fixation and awareness are associated, but measured fixation provides no guarantee of measured awareness, and the converse is also not necessarily true.

This research establishes two important conclusions. First, rates of inattentional blindness vary substantially and are influenced by many factors, including difficulty of the primary task, stimulus properties, and location. Second, rates of inattentional blindness are high; rates in the range of 30% to 60% are not unusual.

Similar studies specifically relevant to aviation displays and in multiple domains show that looking without seeing occurs even by experts in their domain of expertise. In a flight simulator study investigating use of head-up displays in low-visibility landing conditions, experienced pilots encountered an unexpected scenario with a clearly visible aircraft on the runway. Two of the four pilots did not notice the aircraft and proceeded to land despite its presence. These pilots expressed great surprise when this was later pointed out (Fischer et al., 1980). In a study of radiologists detecting lung nodules, 20 of 24 experts failed to detect a large, anomalous figure, and of these, 18 fixated the figure for an average of over half a second (Drew et al., 2013). Further, in actual practice, a post-surgery image was inspected by multiple doctors before a highly visible but unexpected anomaly was detected (Lum et al., 2005). Naval combat pilots showed high levels of change blindness when changes occurred across dynamic screen displays (DiVita et al., 2004; Durlach, 2004).

A complex but informative simulator study with experienced airline pilots investigated awareness of three artificial changes to an autopilot mode (Mumaw et al., 2000). This study measured both eye fixations on key information and the ability to detect the inappropriate mode. Detection could be indicated by either comments or corrective action. Across the three situations, only 1 of 16 pilots (or 1 of 48 opportunities) reported an inappropriate mode. However, eye-tracking data showed that in 42 of these 48 opportunities, the relevant information was fixated, producing a conditional probability of detection given fixation of less than .03. This study also assessed what pilots know about the modes and the context where they should occur. Thus, we can also ask about detection when there is good evidence that the pilot understands the importance of the mode changes being encountered. For example, seven pilots both showed accurate understanding of which mode is appropriate (VNAV SPD vs VNAV PTH) and fixated the inappropriate mode, yet six of these seven provided no indication of any awareness. The one user who did detect it fixated that item ten times over a period of several minutes (see Table 3-10 in Mumaw et al.). This study highlights the complications of investigating what is detected within the flow of complex work domains. Knowledge about the significance and hence need to report can vary and even where knowledge is highly likely, fixation does not ensure awareness.

In summary, while clearly we are often aware of what we are fixating, "looking without seeing" is remarkably pervasive, even by experts in their domain of expertise. People can fixate a

location displaying highly visible information yet be unaware of the existence of the object or the nature of the information.

On the other side of the relationship between fixating and awareness, some studies have reported successful awareness without a measured fixation (Pappas, Fishel, Moss, Hicks, & Leech, 2005), and fixation patterns that do not differ between events where the unexpected object is or is not detected (Beanland & Pammer, 2010; Koivisto, Hyönä, & Revonsuo, 2004). One potential explanation of these results is error in measuring fixation, but it may also be that information can be perceived more peripherally or that the information was inferred rather than observed. The research makes clear that effectively using and interpreting eye-tracking equipment is far from straightforward (Kalar, Liston, Mulligan, Beutter, & Feary, 2016) and there are methodological challenges and limitations in identifying fixation location and duration.

Fixation is neither necessary nor sufficient for awareness, and the degree of association between fixation and awareness varies greatly across situations. For any operational context, we do not know what the strength of association may be.

The objective here was neither to provide a thorough review nor to assess methodological strengths of specific studies. Rather, we wished to highlight that measures of fixation can provide an unknown and perhaps modest amount of information about awareness. This limitation contradicts our intuitions, which inflate the appeal of eye-tracking as a measure of awareness beyond its likely value. If the evaluation goal is determining whether a pilot using a particular display of information is aware of that displayed information, eye-tracking is unlikely to provide a good measure, on its own, for this assessment goal.

Eye-tracking may have more consistent value in evaluating aspects of interface design or user behavior other than awareness. For example, fixation patterns over time and across interface displays may provide information about how attention is distributed *on average* in a particular situation, and how this may change with changes in situation, in user goal, or in display design. Eye-tracking data may provide information about average use or about information-seeking behavior.

## Appendix D: Physiological Measures

Physiological measures do not show the contents of awareness, but a variety of patterns identified by physiological measures do reflect cognitive and affective factors such as workload or stress. Measures include peripheral measures such as skin conductance (also called electrodermal activity, and galvanic skin response GSR), Electrocardiography (ECG), and pupilometry as well as central measures of brain activity such as electroencephalography (EEG), Functional Near-Infrared Spectroscopy (fNIR or fNIRS), and functional magnetic resonance imaging (fMRI). Some measures can be collected with an eye-tracking device (Pupilometry) or with sensors in contact with skin or scalp (skin conductance, ECG, EEG, fNIR), but fMRI requires a prone, isolated position in the scanning device and hence is unlikely to be applicable for device evaluation.

Certain patterns from these measures do show a relation to some psychological variables that might be useful in interface evaluation. Note that all the relations to cognitive variables are moderated by physiological variables, e.g. pupil dilation to variations in task demand is generally much smaller than to variation in level of illumination. Thus, it is important to control or take into account the non-psychological factors, e.g. closely control level of illumination. Examples of these relations include the following:

- Stress or arousal can be indicated by increased skin conductance.
- Attentional demand and workload can be measured by pupil dilation. This has been long established in static tasks (Kahneman & Beatty, 1966; Kahneman, 1973) and more recently for dynamic displays (Iqbal, Zheng, & Bailey, 2004).
- The relation between pupil dilation at a fixation point and target detection has been explored (Vachon & Tremblay, 2014; Vachon et al., 2012) but is not an established measure.
- Target detection and target interpretation influence the EEG pattern labeled P3a and P3b (formerly P300) ERP (event-related potential), which is a most widely used physiological indicator of a moderately specific cognitive process. It is particularly useful for measuring time of detection and of recognizing meaning.
- Increased stress and/or increased workload have been associated with increased heart rate and decreased heart rate variability.

Various levels of support exists for these relations.

In summary, several physiological measures are under investigation, both to identify patterns that predict task-relevant cognitive states, and to assess whether and how such correlates might be used in interface evaluation. Recent reviews have appeared that summarize use of brain measures in aviation and provided broad review of physiological measures with a focus on interface evaluation in HCI (Borghini, Arico, DiFlumeri, & Babiloni, 2017; Cowley et al., 2016).

## Appendix E. Organizational Tool for Assessing Severity of Issues Across Scenarios.

It can be helpful to record and organize the interface issues of concern and the contexts where each issue is likely to affect performance or to be a concern. The results of an Issue-oriented Cognitive Walk-through can be recorded in a matrix marking issues of possible concern and the situations in which the issue is likely to be a problem. The Issue X Scenario Matrix shown here provides a way for tracking the issues as they are identified and linking the issues to situations and tasks planned or used for assessment. The cells in the matrix allow the evaluator to enter an estimate of how problematic the issue is likely to be in that scenario.

Building up and reviewing the matrix allows identification both of the issues most in need of assessment and the situations that provide efficient tests of multiple issues or concerns. For example, a small set of situations and tasks may be informative about multiple issues. Thus, directing inspection or performance-based evaluation to these situations may produce particularly efficient test strategies. In the example below, the "Issue Level-of-concern" column estimates the seriousness of each issue by summing across scenarios and estimates the informativeness of the scenario by summing across each individual issue.

The matrix below lists issues relevant to attention, awareness, and understanding. A broader consideration of evaluation issues is presented in (Mumaw, Haworth, Billman, & Feary, 2019), which also links issues to the related certification rules.

| | ... | Task & Situation: Scenario I | Task & Situation: Scenario J | Task & Situation: Scenario K | Task & Situation: Scenario L | ... | Issue Level-of-concern |
|---|---|---|---|---|---|---|---|
| *1. Provides information the pilot is looking for.* | | | | | | | 0 |
| Needed information is missing. | | | | | | | 0 |
| Perceptual representation of the information is inadequate. | | | | | | | 0 |
| Labels, icons, or messages may be hard to identify or understand. | | | | | | | 0 |
| Context variables may be missing or inadequately displayed. | | | | | | | 0 |
| Display of static reference information may be problematic. | | | | 1 | | | 1 |
| ... | | | | | | | |
| *2. Ensures provided information is accessible and manageable.* | | | | | | | |
| Navigation paths may be unclear, too long, or poorly organized. | | | | 2 | | | 2 |
| Display management methods may be inadequate. | | | | | | | 0 |
| Data validity is not indicated. | | | | | | | 0 |
| Transient information is not adequately preserved. | | | | | | | 0 |
| ... | | | | | | | |

*(continued on next page)*

| | ... | Task & Situation: Scenario I | Task & Situation: Scenario J | Task & Situation: Scenario K | Task & Situation: Scenario L | ... | Issue Level-of-concern |
|---|---|---|---|---|---|---|---|
| *3. Directs pilot attention to information about important changes.* | | | | | | | |
| No alerting scheme was used for an important change. | | 1 | | | | | 2 |
| The alert is not sufficiently salient to draw attention. | | | | | | | 0 |
| The meaning of the alert was not clear. | | | | | | | 0 |
| The interface fails to prioritize and coordinate multiple alerts. | | | | | | | 0 |
| ... | | | | | | | |
| *4. Supports situation understanding and assessment of action in the operational context.* | | | | | | | |
| Interface does not support projecting future events. | | | 3 | 1 | | | 4 |
| Interface does not support fluent updating of an integrated situation model. | | | 3 | | | | 3 |
| The interface does not support accessing or weighting cost/benefit of relevant alternative action choices. | | | 1 | | | | 1 |
| ... | | | | | | | |
| **Informativeness of scenario for awareness and understanding** | | 1 | 7 | 5 | 0 | | |

## References


Beanland, V., & Pammer, K. (2010). Looking without seeing or seeing without looking? Eye movements in sustained inattentional blindness. *Vision Research, 50*(10), 977–988. https://doi.org/10.1016/j.visres.2010.02.024.

Borghini, G., Arico, P., DiFlumeri, G., & Babiloni, F. (2017). *Industrial Neuroscience in Aviation: Evaluation of Mental States in Aviation Personnel.* Cham: Springer.

Burns, C. M., & Hajdukiewicz, J. R. (2004). *Ecological interface design*. Retrieved from http://www.loc.gov/catdir/enhancements/fy0647/2004045492-d.html.

Cowley, B., Filetti, M., Lukander, K., Torniainen, J., Henelius, A., Ahonen, L., … Jacucci, G. (2016). The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human–Computer Interaction. *Foundations and Trends® in Human–Computer Interaction, 9*(3–4), 151–308. https://doi.org/10.1561/1100000065.

DiVita, J., Obermayer, R., Nugent, W., & Linville, J. M. (2004). Verification of the Change Blindness Phenomenon While Managing Critical Events on a Combat Information Display. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(2), 205–218. https://doi.org/10.1518/hfes.46.2.205.37340.

Doane, S. M., Sohn, Y. W., & Jodlowski, M. T. (2004). Pilot Ability to Anticipate the Consequences of Flight Actions as a Function of Expertise. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(1), 92–103. https://doi.org/10.1518/hfes.46.1.92.30386.

Drew, T., Võ, M. L.-H., & Wolfe, J. M. (2013). The Invisible Gorilla Strikes Again: Sustained Inattentional Blindness in Expert Observers. *Psychological Science, 24*(9), 1848–1853. https://doi.org/10.1177/0956797613479386.

Durlach, P. (2004). Change Blindness and Its Implications for Complex Monitoring and Control Systems Design and Operator Training. Human-Computer Interaction, 19(4), 423–451. https://doi.org/10.1207/s15327051hci1904_10.

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*.

Durso, F. T., & Gronlund, S.D. (1999). Situation Awareness. In F. T. Durso & et al. (Eds.), *Handbook or Applied Cognition*. John Wiley & Sons, Inc.

Durso, F. T., & Sethumadhavan, A. (2008). Situation Awareness: Understanding Dynamic Environments. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 442–448. https://doi.org/10.1518/001872008X288448.

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.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. The MIT Press.

FAA: Performance-based operations Aviation Rulemaking Committee (PARC). 2013). *Operational Use of Flight Path Management Systems.*

Fennell, K., Sherry, L., Roberts, Jr., R. J., & Feary, M. (2006). Difficult Access: The Impact of Recall Steps on Flight Management System Errors. *The International Journal of Aviation Psychology, 16*(2), 175–196. https://doi.org/10.1207/s15327108ijap1602_4.

Fischer, E., Haines, R. F., & Price, A. (1980). *Cognitive Issues in Head-up Displays* (NASA Technical Paper No. 1711). Washington, D.C.

Harel, J., Koch, C., & Perona, P. (2007). Graph-Based Visual Saliency. *Neural Information Processing Systems 19.*

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). *Task-evoked pupillary response to mental workload in human-computer interaction.* 1477. https://doi.org/10.1145/985921.986094.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259. https://doi.org/10.1109/34.730558.

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

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science, 154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583.

Kahneman, D.. (1973). *Attention and effort*. Englewood Cliffs: Prentice Hall.

Kalar, D. J., Liston, D., Mulligan, J. B., Beutter, B., & Feary, M. (2016). *Considerations for the Use of Remote Gaze Tracking to Assess Behavior in Flight Simulators* (NASA/TM—2016–219424). Moffett Field, CA: NASA Ames Research Center.

Koivisto, M., Hyönä, J., & Revonsuo, A. (2004). The effects of eye movements, spatial attention, and stimulus features on inattentional blindness. *Vision Research, 44*(27), 3211–3221. https://doi.org/10.1016/j.visres.2004.07.026.

Levin, D. T., Momen, N., Drivdahl, S. B., & Simons, D. J. (2000). Change Blindness Blindness: The Metacognitive Error of Overestimating Change-detection Ability. *Visual Cognition, 7*(1–3), 397–412. https://doi.org/10.1080/135062800394865.

Loft, S., Bowden, V., Braithwaite, J., Morrell, D. B., Huf, S., & Durso, F. T. (2015). Situation Awareness Measures for Simulated Submarine Track Management. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(2), 298–310. https://doi.org/10.1177/0018720814545515.

Lum, T. E., Fairbanks, R. J., Pennington, E. C., & Zwemer, F. L. (2005). Profiles in Patient Safety: Misplaced Femoral Line Guidewire and Multiple Failures to Detect the Foreign Body on Chest Radiography. *Academic Emergency Medicine, 12*(7), 658–662. https://doi.org/10.1197/j.aem.2005.02.014.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, Mass.: MIT Press.

Medina, M., Sherry, L., & Feary, M. (2010). Automation for task analysis of next generation air traffic management systems. *Transportation Research Part C: Emerging Technologies, 18*(6), 921–929. https://doi.org/10.1016/j.trc.2010.03.006.

Most, S. B., Scholl, B. J., Clifford, E. R., & Simons, D. J. (2005). What You See Is What You Set: Sustained Inattentional Blindness and the Capture of Awareness. *Psychological Review*, 112(1), 217–242. https://doi.org/10.1037/0033-295X.112.1.217.

Mumaw, R. J., Billman, D., & Feary, M. (2018). *Factors that Influenced Airplane State Awareness Accidents and Incidents. CAST SE-210 Output 2 Report 2 of 6.* NASA Ames Research Center.

Mumaw, R. J., Billman, D., & Feary, M. (2019). *Identification of Scenarios for System Interface Design Evaluation. CAST SE-210 Output 2 Report 5 of 6.* NASA Ames Research Center.

Mumaw, R. J., Haworth, L., Billman, D., & Feary, M. (2019). *Evaluation Issues for a Flight Deck Interface. CAST SE-210 Output 2 Report 4 of 6.*

Mumaw, R. J., Haworth, L., & Feary, M. (2018). *The Role of Alerting System Failures in Loss of Control Accidents. CAST SE-210 Output 2 Report 3 of 6* [NASA/TM-2019-220176]. NASA Ames Research Center.

Mumaw, R. J., Sarter, N. B., Wickens, C. D., Kimball,S., Nikolic, M., Marsh, R., … Xu, X. (2000). *Analysis of Pilots' Monitoring and Performance on Highly Automated Flight Decks* (Final Project Report No. NASA Ames Contract NAS2).

Neisser, U., & Becklen, R. (1975). Selective looking: Attending to visually specified events. *Cognitive Psychology, 7*, 480–494.

Nielsen, J. (1994). Heuristic Evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods.* New York, N.Y: John Wiley & Sons.

Pappas, J. M., Fishel, S. R., Moss, J. D., Hicks, J. M., & Leech, T. D. (2005). An Eye-Tracking Approach to Inattentional Blindness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 49*(17), 1658–1662. https://doi.org/10.1177/154193120504901734

Parkhurst, D. J., & Niebur, E. (2005). Stimulus-Driven Guidance of Visua Attention in Natural Scenes. In L. Itti & G. Ress (Eds.), *Neurobiology of Attention.* Elsevier.

Pina, P. E., Donmez, B., & Cummings, M. L. (2008). *Selecting Metrics to Evaluate Human Supervisory Control Applications* [HAL2008-04]. Cambridge, MA, US.

Polson, P. G., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies, 36*(5), 741–773. https://doi.org/10.1016/0020-7373(92)90039-N.

Pritchett, A. R., & Hansman, R. J. (2000). Use of testable responses for performance-based measurement of situation awareness. In Endsley, Mica R. & Garland (Eds.), *Situation awareness analysis and measurement* (pp. 189–209).

Sauro, J., & Lewis, J. R. (2012). *Quantifying the User Experience: Practical Statistics for User Reseach.* Retrieved from http://www.123library.org/book_details/?id=46427.

Sherry, L., Feary, M., Polson, P. G., & Fennell, K. (2003). *Drinking from the Fire Hose: Why the Flight Management System Can Be Hard to Train and Difficult to Use* (NASA Technical Memoranda No. 2003–212274).

Sherry, L., Medina, M., Feary, M., & Otiker, J. (2008, May*). Automated tool for task analysis of NextGen automation.* 1–9. https://doi.org/10.1109/ICNSURV.2008.4559185.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*(9), 1059–1074. https://doi.org/10.1068/p2952.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences,* 1(7), 261–267. https://doi.org/10.1016/S1364-6613(97)01080-2.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, *5*(4), 644–649. https://doi.org/10.3758/BF03208840.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences, 9*(1), 16–20. https://doi.org/10.1016/j.tics.2004.11.006.

The Airplane State Awareness Joint Safety Implementation Team (CAST). (2014). *Airplane State Awareness Joint Safety Implementation Team: Final Report Analysis and Recommendations* [Provided to the Commercial Aviation Safety Team].

Vachon, F., & Tremblay, S. (2014). What Eye Tracking Can Reveal about Dynamic Decision-Making. In T. Ahram, Karwowski, W., & Marek, T. (Eds.), *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics*. Kraków, Poland: AHFE.

Vachon, F., Vallières, B. R., Jones, D. M., & Tremblay, S. (2012). Nonexplicit Change Detection in Complex Dynamic Settings: What Eye Movements Reveal. *Human Factors, 54*(6), 996–1007. https://doi.org/10.1177/0018720812443066.

Varakin, D. A., Levin, D., & Fidler, R. (2004). Unseen and Unaware: Implications of Recent Research on Failures of Visual Awareness for Human-Computer Interface Design. *Human-Computer Interaction*, 19(4), 389–422. https://doi.org/10.1207/s15327051hci1904_9.

Wickens, C. D., McCarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., & Zheng, S. (2008). Attention-Situation Awareness (A-SA) Model of Pilot Error. In D. C. Foyle & B. L. Hooey (Eds.), *Human performance modeling in aviation*. CRC Press Taylor & Francis.

Wiggins, S. L., Cox, D. A., & Patterson, E. S. (2010). System Evaluation Using the Cognitive Performance Indicators. In J. E. Miller (Ed.), *Macrocognition Metrics and Scenarios: Design and Evaluation for Real-World Teams.* London: CRC Press.

Wilson, G. F. (2000). Strategies for Psychophysiological Asessment of Situation Awareness. In Endsley, Mica R. & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement.* Mahwah, NJ: Lawrence Erlbaum Associates.