U.S. Department
of Transportation
**Federal Aviation
Administration**

Advisory Circular

MINIMUM FLIGHT CREW

AC 23.1523

JANUARY 12, 2005

Initiated By: ACE-100

FOREWORD

The advisory circular (AC) system became effective in 1962. It provides a single, uniform, agency-wide system that the Federal Aviation Administration (FAA) uses to deliver advisory material to FAA customers, industry, the aviation community, and the public. This AC sets forth one method that may be used to show compliance to the requirements contained within 14 CFR, part 23, § 23.1523, which prescribes the certification requirements for minimum flight crew on part 23 airplanes.

S/

Michael K. Dahl
Acting Manager, Small Airplane Directorate
Aircraft Certification Service

CONTENTS

1. PURPOSE.

a. This advisory circular (AC) sets forth one method that may be used to show compliance to the requirements contained within 14 CFR, part 23, § 23.1523, which prescribes the certification requirements for minimum flight crew on part 23 airplanes.

b. This AC is one method that can be used to determine workload factors and issues for normal, utility, acrobatic and commuter category airplanes. Material in this AC is neither mandatory nor regulatory in nature and does not constitute a regulation.

c. This material is a reference for part 23 airplane manufacturers, modifiers, Federal Aviation Administration (FAA) design evaluation engineers, flight test engineers, engineering flight test pilots [Aircraft Certification Office (ACO), and Flight Standards, and Manufacturers] as well as human factors engineering evaluators. This material may be used by FAA authorized designees to perform workload evaluations.

d. This AC encourages participation and coordination from all the test community participants described in **c**. above. This AC is consistent with the flight test guidance and workload factors described in the minimum flight crew evaluation sections and workload factors described in AC 23-8B. This AC is an acceptable means of showing compliance for part 23 on flight tests and pilot judgments.

2. BACKGROUND.

a. In the early 1980s, there was a move to reduce the crew size of the new generation of commercial jet transport airplanes from three to two. This caused the Transport Airplane Directorate to develop more criteria and guidance for minimum crew determination for part 25 airplanes. AC 25.1523 was developed to provide manufacturers and certification personnel a means of showing compliance to 14 CFR, part 25, § 25.1523. Most part 23 airplanes are single pilot, none require a crew of three, and only a few require a crew of two. There was no need to address crew complement in these airplanes, and no parallel effort was started then for part 23 airplanes. Replacement of conventional electromechanical instruments with complex integrated electronic displays has made it necessary to more closely examine pilot workload for these installations.

b. For many years, part 23 airplane cockpits were rather simple in design and used instruments and systems that were also similar in operation. This made it rather easy for pilots to transition safely from one part 23 airplane to another. However, in recent years because of the growth of modern technology and the reduced cost of electronic components, novel and more complex integrated avionic systems are increasingly being installed in part 23 airplanes. These new systems have changed the appearance, operation, and usability of the pilot-vehicle interface. There is also much variation between manufacturers in the design and operational characteristics of these systems. So, there is a concern that pilot(s) familiar and proficient with one system may not be able to understand sufficiently and operate another system. Although many of these systems can improve pilot situational awareness and safety, poorly designed systems can increase pilot workload, and increase the potential for pilot error. Also, the lack of standardization in the design and operation of these systems can negatively affect pilot training and impact performance and safety. So, there is a need to more closely examine pilot workload and error potential in these complex, integrated cockpits.

c. In August 2002, the Small Airplane Directorate published Human Factors Policy for Part 23 Airplanes. The policy included discussion of the human factors considerations for § 23.1523. The information contained in this AC further describes the human factors considerations referring to that rule. In addition, airplane certification process guidance has also been published to address general certification considerations; these documents include:

(1) The FAA and Industry Guide to Product Certification, 1998

(2) The FAA and Industry Guide to Avionics Approvals, 2001

3. APPLICABILITY. This AC is applicable to all part 23 airplanes as a means of showing compliance to 14 CFR, part 23, § 23.1523, Minimum flight crew. When § 23.1523 was modified by amendment 23-34 in February 1987, additional workload determination criteria were added for commuter category airplanes. The guidance contained

this AC provides recommended evaluation criteria for commuter category airplanes. These considerations may also be used to evaluate workload and usability of the pilot vehicle interface on highly modified cockpits of normal, utility, and aerobatic category airplanes. Historically, the majority of part 23 airplanes have been certified for single pilot operation. It is not expected that this situation will change, as most of the newly developed part 23 airplanes are being designed from the onset to be operated as a single pilot airplane. It is therefore unlikely that the FAA would require two or more crewmembers unless the design was considered to be unsafe for single pilot operation. Therefore, the major focus of this guidance is to provide assistance in examining the acceptability of the workload for single pilot operations. Consequently, for those part 23 airplanes previously certified or designed from the onset for single pilot operation, this AC should be used to examine pilot workload for the purposes of evaluating the adequacy of the cockpit layout, display formats, information presentation, control operation, and system operation logic to support single pilot operation. Generally, the results of these evaluations will be used to identify areas (i.e. as those mentioned in the previous sentence) that may need modification to allow for single pilot operation.

4. RELATED REGULATIONS AND PUBLICATIONS.

a. Regulations: The following related regulations are provided as a quick reference.

23.771 Pilot Compartment
23.773 Pilot Compartment View
23.777 Cockpit Controls
23.1301 Function and Installation
23.1303 Flight and Navigation Instruments
23.1305 Power Plant Instruments
23.1309 Equipment, systems, and installations
23.1311 Electronic display instruments systems
23.1321 Arrangement and visibility
23.1322 Warning, caution, and advisory lights
23.1367 Switches
23.1381 Instrument lights
23.1523 Minimum flight crew
23.1525 Kinds of operation
23.1541 Markings and placards: General
23.1543 Instrument markings: General
23.1545 Airspeed indicator
23.1549 Powerplant and auxiliary power unit instruments
23.1551 Oil quantity indicator
23.1553 Fuel quantity indicator
23.1555 Control markings
23.1583 Operating limitations

b. AC's. Copies of current publications of the AC's listed below can be obtained without charge from the U.S. Department of Transportation, Subsequent Distribution Office, M-30, Ardmore East Business Center, 3341 Q 75th Avenue, Landover, MD 20785. The website where these advisory circulars can be found is http://www.airweb.faa.gov/AC.

AC 23.1311-1A Installation of Electronic Displays in Part 23 Airplanes

c. Policy Statements. You may obtain copies of the current policy statements listed below from the web at http://www.faa.gov/certification/aircraft/ by selecting “Regulations, Policy and Guidance” and then selecting “Policy”. You may request a copy from the Small Airplane Directorate, 901 Locust St., Kansas City, MO 64106.

Part 23 Human Factors Policy PS-ACE100-2001-004

d. AC's for Sale. You may obtain current copies of the following publications from the web at http://www.airweb.faa.gov/AC. If you cannot download them from the web, you may buy them from the

Superintendent of Documents, U.S. Government Printing Office, P. O. Box 371954, Pittsburgh, PA 15250–7954.

AC 23-8B Flight Test Guide
AC 23.1309-1C Equipment Systems and Installations in Part 23 AirplanesDraft

5. GENERAL CERTIFICATION CONSIDERATIONS.

a. Discussions on crew complement and crew workload should take place between the FAA and the applicant early in the program and part of this discussion should address the required level of testing. Minor modifications to the cockpit may not warrant a workload analysis; however, major modifications may require an extensive evaluation. The level of testing should be consistent with the extent of the modifications. The level of required testing should be developed by the applicant and provided to the FAA for concurrence. These discussions should also focus on identification of design features that are likely to impact crew workload. Resulting analyses, demonstrations and tests should be structured to verify that these design features do not place excessive workload demands on any crewmember or increase error likelihood. Usually, the applicant assumes a given crew size at the beginning of the program and there is little likelihood that a different number will be required. In such cases, evaluations of the minimum crew complement must still be conducted to validate the applicant's proposed minimum crew complement. Also, these evaluations are necessary to corroborate the applicant's estimated crew workload by demonstration in a representative operational environment and, thus, support compliance with 14 CFR, part 23, § 23.1523.

b. Minimum crew complement may be established in the beginning of the program considering anticipated pilot-vehicle interface crew workload, individual performance and other suitable factors. Final determination however, will require an evaluation of the fully functioning and integrated system. Because of the direct impact on workload and crew performance, the pilot interface with the cockpit equipment should be evaluated using structured ground tests, mockups and simulators. Applicants are encouraged to include FAA pilots and human factors personnel in these evaluations. We recommend that several flights be conducted in an operationally representative environment. We also recommend the test team be comprised of three to five pilots of varying background and experience to address individual differences in the general pilot population and match the level expected in field service for the aircraft/system under test. A flight test team with diverse backgrounds and experience will expose the system to various piloting techniques and skills. Operationally representative flights can reveal workload excesses that may require workload redistribution or help identify necessary changes in design, operation, training or procedures.

c. Many of the issues related to the human factors design of the cockpit are identified and discussed in the human factors part 23 policy listed in the reference section. Also, previous research and experience have identified several human behaviors that are typically associated with high workload. A list of these behaviors is contained in Appendix 1. It is recommended the test conductor/observer be familiar with the issues discussed in the policy and the list contained in Appendix 1. The flight test evaluator should be vigilant in these areas during testing, noting when and under what conditions any of these behaviors are observed. If any of these behaviors are observed during the evaluations, more detailed examinations of workload may be necessary.

d. When performing analyses and demonstrating workload for minimum crew determinations, several workload factors have been identified as significant in performing this task. These workload factors are listed in Appendix 2 and need to be addressed as part of the evaluations. It is important to include all of the workload factors listed in the appendix when showing compliance to the minimum flight crew rule 14 CFR part 23, § 23.1523.

6. CERTIFICATION PROCEDURES.

a. Evaluation Test Plan. Any new or extensively modified airplane should be evaluated using an FAA approved test plan. Methods for substantiating compliance with § 23.1523 may include the use of analyses, simulation, demonstrations, and/or flight tests. The minimum crew complement workload should be examined through a logical process of analysis, measurement, and demonstration of the workload encountered in the new or modified cockpit.

b. Analytical Approach.

(1) General. The applicant should conduct analyses at the beginning of the design process. Methods of analysis should be selected on the basis of predictive validity, reliability, and applicability to the particular cockpit configuration. When possible, the analyses should include comparisons with other cockpits similar to the cockpit under consideration. The focus of any analysis should be on the modifications or new equipment. More significant modifications, which largely change the cockpit configuration, pilot-vehicle interface, pilot tasking or operating procedures, will require a more comprehensive evaluation. Applicants are encouraged to share the results of these analyses with the certifying office as soon as possible to identify and resolve potential issues.

(2) Examples. One acceptable analytical approach assesses workload as a percentage of the time available to perform tasks (Time-Line Analysis) (other acceptable analytical approaches are described in Appendix 3). This process should be applied to an appropriate set of flight segments in which operationally important time constraints can be identified. This method is satisfactory for evaluation of cockpit changes about overt pilot tasks such as control movements and data inputs. The accepted practice involves careful selection of a limited set of flight scenarios and time segments that represent the range of operational requirements (including the range of selected normal, nonnormal, and emergency procedures). Task time-line analysis yields useful data when tasks must be performed within significant time constraints. An accurate determination of the time available is critical if this method is to have any value. Although time measurements cannot be interpreted to be an absolute standard, such data can be used to identify increased workload demands, which may require a more thorough examination. The impact of cockpit changes on tasks involved with planning and performing emergency or nonnormal procedures should be closely examined. If preliminary analyses by the certification team identify any potential problem areas, a more extensive evaluation should be conducted.

c. Testing.

(1) Overview. Testing to substantiate crew complement and to evaluate crew workload should be an iterative process beginning early in the program. The process should include tests in mockups, bench setups, or simulators and should progress to flight tests in the target airplane with the final configuration. It is useful and/or necessary to incorporate simulator-based testing during the development process to obtain preliminary performance evaluations before flying the prototype system in the airplane. Although not required, it is worthwhile to conduct some evaluations using participants other than company or FAA test pilots (i.e., individuals who are not members of the manufacturer's or FAA certification team). If used, these individuals should be representative of the target population for experience and training, which would be appropriate and expected for the system/airplane. The final determination of the minimum crew complement is to be reserved until the airplane has been flown by pilots with experience appropriate to the aircraft being certified. Paragraph 4 below contains the criteria for determining the minimum flightcrew under § 23.1523.

(2) Scope. The test program should address all workload functions and factors listed in § 23.1523 and Appendix 2. For example, an evaluation of workload should include the communications tasks required to operate properly the airplane in the environment for which approval is sought. The goal is to evaluate workload with the proposed crew complement during realistic operating conditions, including representative air traffic, weather, airline operational duties, and cabin communications.

(3) Evaluation Methodology. There are mainly three workload methods for conducting these evaluations, *direct comparison, indirect comparison* and *standalone evaluation.* These are all described below. Comparative evaluations (direct or indirect) examine pilot-system performance and acceptance of the new design with a previously certified similar system that has a demonstrated safety and service history in operational service. Typically, if the workload for the new system is rated equal to or less than that of the referenced system, then it can be assumed that workload associated with use of the new system is acceptable. Comparative evaluations may be done directly whereby two systems are evaluated against one another for a given set of tasks and flight segments. Evaluations may also be done indirectly in which only the new system is evaluated but test participants are asked to compare its performance based on their recollection of another system’s performance. When conducting

comparative evaluations, caution should be exercised, as it is possible the previously certified system may have function/operation problems or other undesirable characteristics that would not serve as "good" comparative references. For any comparative evaluation, a thorough review of the safety and service history should be conducted to ensure there are no problems or issues with the comparison system. The other method of evaluation is to evaluate a system as a stand-alone component, where no comparison is made to any other system, but rather the system is evaluated based on its own merits. However, the new component must be evaluated as part of the overall cockpit system, since integrating this component may affect task performance and workload of other systems in the cockpit. The stand-alone approach requires establishment of success criteria for each data collection tool that will be used for the evaluation before the start of testing. For example, if the Bedford Workload Scale is used to collect data on subjective workload, the evaluator should show the scale value at which the workload is considered unacceptable before testing begins. Although collected data may show the workload associated with a given task is unacceptably high, the final determination of whether workload and/or performance warrants changes in design, procedures, and/or training should made by the test team. For items in which only one or two test team members experienced a problem, the team should factor in the skill level, training, learning ability and experience of those individuals. This ensures the differences were not the result of deficiencies in one or more of these areas.

(a) Direct Comparisons. For this approach, the preferred method for conducting a comparative evaluation is to structure tests such that data are collected from evaluation participants performing comparable tasks with both the new design and the previously approved like system. Specific evaluations should be conducted using scenarios designed to exercise fully the new design features. It is important that tasks, to be performed with both systems, be structured and scripted to be as exact and repeatable as possible to promote direct comparisons of pilot-system performance between the two systems. Also, the interval between the evaluations of the two systems should be kept to a minimum.

(b) Indirect Comparisons. Occasionally, it may not be possible to conduct the multiple evaluations necessary to compare directly one system against another. It is acceptable in such cases to "indirectly" compare the systems by requiring test participants to "mentally" (i.e., based on their memory of a reference system performance and workload) compare their performance with the new design/system against a previously approved like system. Care should be taken when analyzing results from such comparisons, as the data may not be as accurate or reliable as direct comparison data. So, the conclusions drawn from such testing should be reviewed carefully, checking for reasonableness and consistency.

(c) Standalone Evaluations. This method may be used when a direct or indirect comparison with a reference (baseline) system may not be suitable, if the new design differs radically from the other currently certified systems. For example, if the new design represents a significant change in the level of automation or pilot duties, comparison with a reference design may have limited value. In this case, there is a lack of a baseline for comparison; therefore, the evaluation of the system must be based on its own merits and intended function that has been determined by the applicant and agreed to by the FAA. Evaluation tasks should be well described in adequate detail to exercise fully system functions, implied by their intended function, under all probable operational conditions.

(4) Scenarios. Evaluations should be structured to assure that new systems and cockpit configurations are evaluated using scenarios representative of the kinds of operation for which the airplane is authorized. Although the applicant may provide quantitative substantiating crew workload data, structured subjective evaluations are usually required to verify submitted data in operationally representative environments. These evaluations compare the ease of execution of crew tasks in the subject airplane under identical or substantively similar scenarios. A proposed flight test program for showing compliance with § 23.1523 should address the following factors:

(a) Routes. The test program routes should be modelled to provide a representative mix of navigation aids, airports, instrument approaches and Air Traffic Control (ATC) services.

(b) Weather. The routes should be selected to provide the likelihood of encountering types of adverse weather appropriate to the airplane's intended operation [Instrument Meteorological Conditions (IMC) conditions, night, turbulence, icing, etc.].

(c) Crew Work Schedule. The test crew should be assigned to a daily work schedule that is representative of the operations for which the airplane was developed. The program should include the duration of the workday and the maximum expected number of departures and arrivals, flights that begin at night, maximum allowable duty times, and minimum rest periods.

(d) Minimum Equipment List. The applicant should incorporate representative dispatch configurations in the proposed flight test program. Combinations of these representative dispatch configurations with probable resulting simulated malfunctions should form the basis of many of the evaluation scenarios.

(e) Intended Operations. The airplane should be operated on routes that would adequately sample high-density areas in both IMC and Visual Meteorological Conditions (VMC). These routes would include the types of approaches expected in the intended operations, such as precision/non-precision approaches, missed approaches, holding and diversions to alternate airports.

(f) Incapacitated Crewmember. Most part 23 airplanes are certified for single pilot operations, however there are some exceptions in which two crewmembers are required. In those cases, a workload evaluation for an incapacitated crewmember should be conducted. The certification program should include a demonstration of operations during the total incapacitation of a crewmember at any point in a given flight. It must be shown the airplane can safely operate and land with the remaining crew. In the event of an incapacitated crewmember, the pilot would notify ATC and special considerations would be given to aid the pilot in the approach and landing to a given airport. Therefore, it is not necessary to require performing this demonstration to a “high” density airport, as it would not accurately represent task requirements and demands. Incapacitated crewmember tests need not be additive to all other "dispatch plus subsequent failure" scenarios. Incapacitation should be viewed as another example of "subsequent failure" to be included within one or more scenarios beginning with a dispatch configuration, which includes selected items from the proposed minimum equipment list.

(g) System Failures. The workload outcomes of system failures and degraded modes of operation should be evaluated in the program. Failures of both primary and secondary systems and a representative combination of failures should be included. Data should be collected on pilot recognition time, interpretation accuracy, and the appropriateness of resulting actions to acknowledge, correct and/or compensate for the failure. A sampling of various emergencies and nonnormal conditions should be set up in the test program to show their effect on the crew workload. Data should be collected on the pilot's ability to recognize and take suitable action to emergencies and nonnormal situations. During these evaluations, it is a good time to verify the adequacy of associated checklists and procedures.

NOTE: Before selecting the system failures that will be evaluated in the flight test program, simulation or analytical studies should be conducted to determine which failures are more probable. System failures should be selected based on the likelihood of occurrence. It is not necessary to evaluate those failures that have a low chance of happening (i.e. 1×10^{-9}). The crew workload distribution, during the executing of emergency or nonnormal situations, should be understood to assure selection of appropriate failure cases. Guidelines implementing a selected number of subjective, physiological, and performance workload measurement techniques, are contained in the FAA sponsored report "Assessment of Crew Workload Measurement Methods, Techniques, and Procedures" Vol. II (Report No. WRDC-TR-89-7006).

(h) Error Evaluation. The cockpit design should be evaluated to determine if workload is a significant contributor to pilot errors. Control and display design deficiencies could significantly contribute to pilot error. An example of evaluation of this type can be found in Appendix 3 under workload and error (reference Test Operation Procedure 1-2-610, test procedure-error likelihood analysis). An examination of errors should also be made to ensure that when they occur they do not create an unsafe or hazardous condition. Workload and error are discussed further in Appendix 3 with related references.

d. Data Collection and Analysis.

(1) Workload Data. Data may contain objective performance measurements, subjective (perceived) performance estimates collected through questionnaires and pilot comments, and observations recorded by the test conductor, flight test engineer or human factors engineer. Typically, objective data (time-to-perform a task, error

rates, number of separate actions, etc.) are preferred to subjective data. However, much can be learned from subjective experiences, observations and opinions. Also, there may be cases when it may be difficult to collect precise objective data because of test setup limitations or other variables that cannot be controlled nor compensated for by the test team. Under such circumstances, subjective data alone may be used to show compliance to 14 CFR, part 23, § 23.1523. For evaluations where subjective data only is collected, the applicant should consider the qualitative rankings of the perceived workload factors listed in Appendix 2 as well as a class of observable pilot behaviors listed in Appendix 1. It is recommended, however, that a combination of all the methods (both objective and subjective) be employed. Each combination can provide valuable information about the adequacy of the interface, workload and the ability of pilots to understand and use the system. It is the responsibility of the certification authority to determine whether a system successfully meets this requirement. The determination should be based on a review of all data collected throughout the test program, with particular emphasis on the workload, duration of the task, uncompleted tasks, and error rates.

(2) Forms. Data collection forms, questionnaire development, rating scales and other test reporting details are described in Appendix 4, Recommended Test Plan Guidance and Data Collection Information.

e. Credit for Tests. Applicants should conduct cockpit design and pilot performance evaluations throughout the development of a new system or component. The FAA may grant credit for performing such tests if it can be shown the test was conducted on a conformed component/system. The conformity may be done following the test, although the applicant must substantiate the component or system used for the evaluation is the same conformed component/system. Well-conducted and documented evaluations of a system early in a program may reduce flight-testing and FAA involvement later in the certification program. Credit for such evaluations must be coordinated before such testing with the responsible FAA office or authorized representative.

APPENDIX 1. PILOT BEHAVIOR ASSOCIATED WITH HIGH WORKLOAD

A number of human behaviors that are typically associated with high workload and they are listed below. The test conductor/observer should be especially cognizant of these behaviors during the conduct of the test and pay particular attention as to when and under what conditions any of these behaviors are observed. If any of these behaviors are observed during the course of the evaluations, more detailed examinations of workload may be necessary.

1. Observed learning behaviors

a. Pilot expends excessive effort to learn use of equipment

(1) Requires substantial assistance

(2) Repeatedly refers to training material

b. Pilot requires extended time period to become familiar with operation

c. Memory requirements are excessive

(1) Makes repeated reference to manuals throughout training for procedures

(2) Difficulty finding functions within menus

(3) Difficulty in returning to initial display page, mode, etc.

2. Observed cockpit operational behavior

a. Repeatedly refers to manuals or handbooks

b. Makes repeated unsuccessful attempts to enter or alter data

c. Exhibits excessive head-down time

(1) Break-down of normal scan

(2) Fixation on single display or interface

(3) Insufficient monitoring of surrounding airspace in VMC

d. Insufficient monitoring of cockpit displays

e. Performs task shedding

(1) Multiple concurrent tasks are not all performed (i.e., fails to perform checklist items)

(2) Overuse of automation to maintain performance

(3) Communications tasks

(a) Responses are monosyllabic

(b) Not all communications are acknowledged

(c) Read-backs are incorrect

APPENDIX 1. PILOT BEHAVIOR ASSOCIATED WITH HIGH WORKLOAD (CONTINUED)

(d) Pilot misses communications

f. Commits errors

(1) Discrete

(a) Errors of commission (e.g., selects incorrect mode)

(b) Errors of omission (e.g., fails to engage mode when required)

(2) Continuous

(a) Course tracking errors

1. Vertical (altitude)

2. Horizontal (lateral and/or overshoot errors)

g. Exhibits signs of stress or fatigue

(1) Unable to respond to critical items in a timely fashion

(2) Indecisive (prolongs decisions or is unable to make them)

(3) Inattentive (fails to respond to events)

(4) Consistently asks for verbal communications to be repeated

(5) Unfocused (e.g., unable to concentrate on a single task to completion)

APPENDIX 2. MINIMUM FLIGHT CREW RULE

The following restates the contents of 14 CFR part 23, § 23.1523 regarding workload determination:

The minimum flight crew must be established so that it is sufficient for safe operations considering –

(a) The workload of individual crewmembers, and in addition, for commuter category airplanes, each crewmember workload determination must consider the following:

(1) Flight path control.
(2) Collision avoidance.
(3) Navigation.
(4) Communications.
(5) Operation and monitoring of all essential airplane systems,
(6) Command decisions, and
(7) The accessibility and ease of operation of necessary controls by the appropriate crewmember during normal and emergency operations when at the crewmember flight station.

(b) The accessibility and ease of operation of necessary controls by the appropriate crewmember and;

(c) The kinds of operation authorized under 23.1523.

WORKLOAD FACTORS (reference AC 23-8B, Flight Test Guide for Part 23 Airplanes)

The following is a listing of recognized workload factors considered significant when analyzing and demonstrating workload for minimum flight crew determination:

1. The impact of basic airplane flight characteristics on stability and ease of flight path control. Some factors such as trimmability, coupling, response to turbulence, damping characteristics, control breakout forces and control force gradients should be considered in assessing suitability of flight path control. The essential elements are the physical effort, mental effort and time required to track and analyze flight path control features, and the interaction with other workload functions.

2. The accessibility, ease, and simplicity of operation of all necessary flight, power, and equipment controls, including emergency fuel shutoff valves, electrical controls, electronic controls, pressurization system controls, and engine controls.

3. The accessibility and conspicuity of all necessary instruments and failure warning devices such as fire warning, electrical system malfunction, and other failure or caution indicators. The extent to which such instruments or devices direct the proper corrective action is also considered.

4. The complexity and difficulty of operation of the fuel system, with particular consideration given to the required fuel management schedule necessitated by e.g. structural, or other airworthiness considerations. Additionally, the ability of each engine to operate continuously from a single tank or source that is automatically replenished from other tanks if the total fuel supply is stored in more than one tank.

5. The degree and duration of concentrated mental and physical effort involved in normal operation and in diagnosing and coping with malfunctions and emergencies, including accomplishment of checklist, and location and accessibility of switches and valves.

6. The extent of required monitoring of the fuel, hydraulic, pressurization, electrical, electronic, deicing, and other systems while en route and recording of engine readings, and so forth.

APPENDIX 2. MINIMUM FLIGHT CREW RULE (CONTINUED)

7. The degree of automation provided in the event of a failure or malfunction in any of the aircraft systems. Such automation should ensure continuous operation of the system by providing automatic crossover or isolation of difficulties and minimize the need for flight crew action.

8. The communications and navigation workload.

9. The possibility of increased workload associated with any emergency that may lead to other emergencies.

10. Passenger problems.

11. Incapacitation of a flight crewmember whenever the applicable operating rule requires a minimum flight crew of at least two pilots.

APPENDIX 3. WORKLOAD EVALUATION ANNOTATED REFERENCES

1. PURPOSE. The purpose of this appendix is to provide a definition of workload and discuss the relationship of workload and performance. The remainder of the appendix provides an annotated bibliography of the literature on this topic.

a. Definition of Workload. Workload is a term used to describe the relationship between an individual's capacity to perform a task (mental and/or physical), and the level of system and situational demands associated with the performance of that task. The basic notion is related to the differences between the amount of resources available in the human operator and the amount of resources demanded by the task. Tasks that demand much of the human resources (capacity) are considered high workload tasks. Conversely, tasks that demand little of the human resources (capacity) are considered low workload tasks.

b. Performance and Workload. The relationship between workload and performance generally follows an inverted U-shaped curve, such that performance suffers if workload is either too high or too low. Levels of workload that are too low are often found in fully automated systems where the operator serves largely as a monitor of the automated processes. In these cases the operator/pilot may become inattentive and/or bored, and this situation is generally referred to as task underload. At the opposite extreme, levels of workload that are too high often cause the pilot to miss important information, fail to perform tasks, make errors or engage in task shedding in an attempt to reduce workload. Although automation is generally employed to and thought to decrease pilot workload, in some cases it has either had little effect or actually increased workload. While the automation may relieve the pilot from the performance of the automated tasks, it may impose additional task requirements related to operating the automated functions. One difficulty found in highly automated systems is that when the automation fails, pilot workload can be significantly increased. Specific attention should be given to the relative ease or difficulty with which the pilot can detect changes in or extract information about system status.

A key goal of the system designer is to find the middle ground such that the pilot is sufficiently engaged in the process to remain alert and involved in system operation yet not so loaded as to be incapable of performing all of the necessary tasks in an efficient and timely fashion. Additionally, the analysis of workload must be extended to include the case where automation fails, and the pilot must now perform those functions. Therefore, the introduction of new integrated complex avionics systems, whether highly automated or not, needs to be evaluated and should also include an examination of human error.

c. Workload and Error. Human error is often related to workload, and there is usually a positive correlation between excessive workload and the occurrence of errors. It should be noted that errors could also be associated with low workload. When recording flight crew error, please refer to list of annotated references on this topic. The references describe methods and data collection techniques that can be used in the evaluation.

When collecting data the pilot's performance should be quantified in terms of:

(1) Number of errors committed

(2) Consequences of errors

(3) Time spent recovering from errors

(4) Number of subsequent incorrect actions due to the error

(5) Time spent using manuals to recover from errors.

APPENDIX 3. WORKLOAD EVALUATION ANNOTATED REFERENCES (CONTINUED)

Annotated References

Selecting Performance Measures

1. Kantowitz, B.H. (1992). Selecting measures for human factors research. *Human Factors, 34*(4), 387-398.

Selecting measures is a necessary component of human factors research. Proper selection must take into account the representation problem (how is the assignment of numbers to objects or phenomena justified?) and the uniqueness problem (to what degree is this assignment unique?). Other key human factors measurement issues include subject representativeness, variable representativeness, and setting representativeness. It is difficult to create a single measure that captures essential characteristics of complex systems. Several examples illustrate how theory can guide measurement selection in such diverse human factors research as vigilance, turning off warning alarms, information requirements for military command centers, subjective workload, heart-rate signal analysis, and heat stress in nuclear power plants.

Measurement and Assessment Techniques

1. Assessment of crew workload measurement methods, techniques and procedures.
Volume 1, Process, methods and procedures; Volume 2, guidelines for the use of workload assessment techniques in aircraft certification.

Volume 1 states that the goal of this document is to "identify assessment techniques which demonstrate evidence of validity and reliability and are suitable as measures of flightcrew workload for certification. This document was sponsored by the FAA and was a two-year effort. Testing was done via simulation and used various measures including SWAT, NASA TLX, and Bedford rating scales. Timeline analysis was used to identify high and low task demand levels. Volume 2 contains basic information on how to gather workload data.

2. Casali, J.G. and Wierwille, W.W. (1983). A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. Human Factors, 25(6), 623-641.

Sixteen potential metrics of pilot mental workload were investigated regarding their sensitivity to communication load and their intrusion on primary task performance. A moving-base simulator was used to present three cross-country flights. The flights varied only in the difficulty of the communications requirements. Results indicated that both the Modified Cooper-Harper Scale and the workload Multi-descriptor Scale were sensitive to changes in communications load. The secondary-task measure of time estimation and the physiological measure of pupil diameter were also sensitive. As expected, those primary-task measures that were the direct measure of communicative performance were also sensitive to load, whereas aircraft control primary-task measures were not, attesting to the task specificity of such measures.

3. Gawron, V.J., Schiflett, S.G., and Miller, J.C. (1989). Measures of in-flight workload. In R.S. Jensen (Ed.) *Aviation Psychology*. Aldershot, Haunts, UK: Gower Publishing Company.

This reference contains a review of empirical measures of workload that can be used during flight test. The review includes a general description of the workload measures, associated strengths and weaknesses, and guidelines for use. The chapter also includes a discussion of workload validation studies.

4. Gawron, Valerie J. (2000). *Human Performance Measures Handbook*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Drawing from her work testing and evaluating projects of military and commercial transportation systems, Gawron offers help to both the novice and the experienced in selecting measures to be used to evaluate

APPENDIX 3. WORKLOAD EVALUATION ANNOTATED REFERENCES (CONTINUED)

human/machine systems. The handbook begins with an overview of the steps involved in developing a test to measure human performance, workload, and/or situational awareness. This is followed by a definition of human performance and a review of human performance measures. It is a unique reference covering the basics of measures as well as authoritative summaries of over 100 performance, workload, and situational awareness measures. It is an excellent desktop reference for designers, test and evaluation specialists, and system specification developers.

5. Gopher, D., and Donchin, E. (1986). Workload--An examination of the concept. In K. Boff and L. Kauffman (Ed.s), *Handbook of perception and human performance*. New York: Wiley and Sons, chapter 41, pp. 1-49.

This chapter represents a theoretical examination of the multidimensional, multifaceted concept of workload. Due to the complexity of the construct, no single measure is capable of capturing all relevant aspects, nor may multiple measures covary within a single task. The discussion was concerned with clarifying the nature of the dimensions along which workload varies to explicate the attributes that should be considered in the selection of a measurement procedure. The primary thesis is that workload assessment focuses on measuring the processing and response limitations of the human information processing system which are revealed through the interactions between an operator and the assigned tasks. The nature of the limitations were considered on two levels: (1) the more theoretical level (in which the invariant, open loop properties of the human processing system were examined), and (2) a more practical level (in which workload was characterized, at any instant, as the joint, closed loop property of the human and the assigned task). In general, the focus of the theoretical discussions emphasized the close affinity between the study of workload and attention, with an additional discussion of the energetical and structural characteristics of the central processor. The recommendation was made that measurement procedures should encompass both conscious and non-conscious processing activities; a detailed task analysis should be performed to uncover the major components of the task, followed by a battery of performance-based measures designed to evaluate the load on each component.

6. Hart, S. G. (1986). Theory and measurement of human workload. In J. Zeidner (Ed.), *Human productivity enhancement: Training and human factors in system design* (pp. 396-456). New York: Praeger.

The goal of this reference is to define human workload, what influences it, how it is measured, and why it is of theoretical and practical concern. The first section reviews typical definitions and motives for measuring and predicting workload. A structure is proposed to relate and integrate many of the factors that create or influence it (e.g., the experiences of operators). A third section describes five types of assessment and predictive methodologies: (1) subjective ratings, (2) primary task performance, (3) secondary task performance, (4) physiological recordings, and (5) analytic procedures. Finally, the selection and application of appropriate tools to predict or assess imposed workload, system performance and behavior, or operator experience is considered.

7. Hart, S. G. (1987). Research papers and publications (1981-1987): Workload research program (NASA Technical Memorandum 100016). Washington, DC: National Aeronautics and Space Administration.

This document contains an annotated bibliography of the research reports written by participants in NASA's Workload Research Program from 1981 to 1987. It represents the results of theoretical and applied research conducted at Ames Research Center and at universities and industrial laboratories funded by the program. The major program elements include: (1) developing a fundamental understanding of the concept of workload, (2) providing valid, reliable, and practical measures of workload, and (3) creating a computer model to predict workload. The overall goal is to provide workload-related design principles, measures, guidelines, and computational models. The research results are transferred to user groups by establishing close ties with manufacturers, civil and military operators of aerospace systems, and regulatory agencies; publishing scientific articles; participating in and sponsoring workshops and symposia; providing information, guidelines, and computer models; and contributing to the formulation of standards. In

APPENDIX 3. WORKLOAD EVALUATION ANNOTATED REFERENCES (CONTINUED)

addition, the methods and theories that have been developed have been applied to specific operation and design problems at the request of a number of industry and government agencies.

8. Hart, S.G., and Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock, and N. Meshkati (Eds), *Human Mental Workload,* 139-183. North Holland, Amsterdam: Elsevier.

Abstract: The results of a multi-year research programme to identify the factors associated with variations in subjective workload within and between different types of tasks are reviewed. Subjective evaluations of 10 workload-related factors were obtained from 16 different experiments. The experimental tasks included simple cognitive and manual control tasks, complex laboratory and supervisory control tasks, and aircraft simulation. Task, behaviour and subject-related correlates of subjective workload experiences varied as a function of difficulty manipulations within experiments, different sources of workload between experiments, and individual differences in workload definition. A multi-dimensional rating scale is proposed in which information about the magnitude and sources of six workload-related factors are combined to derive a sensitive and reliable estimate of workload.

9. Kantowitz, B.H., and Casper, P.A. (1988). *Human Workload in Aviation*. In E.L. Wiener and D.C. Nagel (Eds.), Human Factors in Aviation. San Diego: Academic Press, Inc.

In Kantowitz and Casper's chapter on Human Workload in Aviation a wide variety of approaches to measuring workload are described, and empirical studies using these methods are reviewed. The importance of the concept of workload and the challenges of defining this hypothetical concept are reviewed. The authors relate workload to the psychological concept of attention, and suggest how theoretical models of attention can provide insight to the problem of measuring human workload in air traffic control and cockpit applications.

10. National Test Pilot School course “human factors and workload” (minimum flight crew)

This course is taught at the National Test Pilot School in Mojave, CA. It was developed in the early 80's as a result of changes to the minimum flight crew rule and development of AC 25.1523 for Part 25 operations. This course was intended to raise the level of awareness and the importance of workload evaluations that were being conducted by the certification test pilot community. FAA and non-FAA evaluators receive this training and instruction materials when they attend the course.

11. Ogden, G.D., Levine, J.M., and Eisner, E.J. (1979). Measurement of Workload by Secondary Tasks. *Human Factors, 21*(5), 529-548.

The post-1965 literature on the use of secondary tasks in the assessment of operator workload was surveyed. Twelve classes of tasks were identified; the most frequently used were choice reaction time, memory, monitoring, and tracking. The literature review did not suggest a single best task or class of tasks for the measurement of workload. Limitations in using secondary tasks are discussed, and directions for future research are presented.

12. Rehmann, J.T., Stein, E.S. and Rosenberg, B.L. (1983). Subjective pilot workload assessment. *Human Factors, 25*(3), 297-307.

Compared workload ratings done after the simulator flight with those done periodically during the flight. An auditory prompt was sounded once per minute and the pilot was required to enter a workload rating on a scale of one through ten using a numeric keypad. They found that ratings done during the flight most accurately reflected the difficulty of that flight and that the latency of pilot response to the auditory prompt was positively correlated with the difficulty of the task (responses were slower when workload was rated as high).

APPENDIX 3. WORKLOAD EVALUATION ANNOTATED REFERENCES (CONTINUED)

13. Sirevaag, E., Kramer, A., Wickens, C., Reisweber, M., Strayer, D., and Grenell, J. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics, 36,* 1121-1140.

This research examined the processing demands imposed upon experienced pilots by two different communication formats, digital and verbal, in a high fidelity simulation of an advanced multi-function helicopter. The mental workload imposed by the type and magnitude of communications was assessed by a battery of subjective, performance, secondary, and physiological measures. The results are discussed in terms of the structural and capacity demands of the communications systems that were proposed for the advanced multi-function helicopter.

14. Vidulich, M.A. and Tsang, P.S. (1986). Techniques of subjective workload assessment: A comparison of SWAT and the NASA Bipolar Method. *Ergonomics, 29* (11), 1385-1398.

Two techniques for assessing subjective workload, the Subjective Workload Assessment Technique (SWAT) and the NASA Bipolar method, were applied in a laboratory experiment that required the rating of a number of single- and dual-tracking and spatial transformation tasks. Both subjective assessment techniques displayed similar sensitivity to the different task manipulations. SWAT assesses subjective workload in terms of time load, mental effort load and psychological stress load. The NASA Bipolar method assesses subjective workload in terms of task difficulty, time pressure, performance, mental/sensory effort, physical effort, frustration, stress, fatigue and activity type. The purpose of the study was to compare the merits of the two techniques.

15. Vidulich, M.A., Ward, G.F., and Schueren, J. (1991). Using the Subjective Workload Dominance (SWORD) Technique for projective workload assessment. *Human Factors, 33*(6) 677-691.

This study examined the utility of the Subjective Workload Dominance (SWORD) technique as a projective workload tool. Two groups predicted the workload associated with using six possible head-up display formats. The projective ratings from the groups were correlated with retrospective ratings from another group that had experienced all six formats in a simulator study. The results support the utility of the SWORD technique as a projective tool, provided a group of subject matter experts is available to make the required judgments.

16. Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors, 34*(4), 457-468.

(a) 80 percent of the usability problems are detected with 4 or 5 subjects, (b) additional subjects are less and less likely to reveal information, and (c) the most severe usability problems are likely to have been detected in the first few subjects. Formulas are provided for calculating the number of subjects needed for a given level of confidence.

17. Williges, R.C. and Weirwille, W.W. (1979). Behavioral measures of aircrew mental workload. Human Factors, *21,* 549-574.

Behavioral research literature pertaining to the measurement of aircrew workload was classified into general categories of subjective opinion, spare mental capacity, and primary task metrics. Fourteen specific classes of workload measures related to these general categories were reviewed specifically with regard to aircrew workload assessment. Each class of measures was summarized in terms of background, applications, and implications for research and implementation. Due to the multidimensionality of workload, it was concluded that the most promising assessment procedure should include multiple measures of subjective opinions, spare mental capacity, and primary task measures as well as physiological correlates.

APPENDIX 3. WORKLOAD EVALUATION ANNOTATED REFERENCES (CONTINUED)

18. Wierwille, W.W. and Conner, S.A. (1983). Evaluation of 20 workload measures using a psychomotor task in a moving-base aircraft simulator. *Human Factors, 25*(1), 1-16

Examined 20 workload measures using a psychomotor task with 3 levels of psychomotor load. Sensitivities of workload techniques vary and only a few of the techniques tested appeared to be sensitive to psychomotor load. Provides suggestions for how to select a workload estimation technique for specific tasks. Used instrument approach and landing as task. Looked at these categories of assessment techniques: opinion, spare mental capacity, physiological measures, and eye behavior. Primary task performance indicated which techniques were most sensitive to workload manipulations.

19. Wierwille, W.W., Rahimi, M., and Casali, J.G. (1985). Evaluation of 16 measures of mental workload using a simulated flight task emphasizing mediational activity. Human Factors, 27(5), 489-502.

Presented evaluation of 16 mental workload measures applied to mediational piloting tasks (defined as activities related to logic, reasoning, decision-making and judgment). Identifies which measures were sensitive to mediational workload, which ones were not, and which measures were intrusive.

Comparison of Workload Measures

1. Hankins, T.C. and Wilson, G.F. (1998). A comparison of heart rate, eye activity, EEG, and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine. 69*(4), 360-367.

Cardiac, eye, brain and subjective data were collected during an actual flight scenario designed to present pilot subjects with tasks requiring different piloting skills at several levels of mental workload. Results from these methods were compared and suggest that while subjective estimates provide useful information about pilot workload, multiple measures, especially psychophysiological measures (cardiac, respiratory, eye, brain and hormonal measures), provide a more comprehensive picture of the mental demands of flight. The measures used in this study provided unique, non-overlapping information.

2. Hicks, T.G. and Wierwillie, W.W. (1979). Comparison of five mental workload assessment procedures in a moving-base driving simulator. *Human Factors, 21*(2), 129-143.

Examined five methods of measuring workload in a driving simulator (a psychomotor task): secondary tasks, occlusion techniques (forced time sharing by limiting the time that visual information could be viewed), physiological measures (cardiac arrhythmia), subjective opinion method, and primary task performance. Primary task and rating scale appeared to be the most sensitive for this application.

3. Hill, S.G., Iavecchia, H.P., Bittner, A.C. Jr., Zaklad, A.L., and Christ, R.E. (1992). Comparison of four subjective workload rating scales. *Human Factors, 34*(4), 429-439.

Four subjective workload scales were compared along four dimensions: sensitivity, operator acceptance, resource requirements, and special procedures. The scales were the Modified Cooper-Harper scale, the National Aeronautics and Space Administration Task Load Index (TLX), the Overall Workload (OW) scale, and the Subjective Workload Assessment Technique. Three U.S. Army systems were studied for potential workload concerns. Data from five different studies on the three systems were compared along the aforementioned four dimensions. Results indicate that all four scales are acceptable tools and are sensitive to different levels of workload. However, TLX and OW are consistently superior when considering sensitivity, as measured by factor validity and operator acceptance. TLX and SWAT may help identify the causes of excessive workload and ways to alleviate it.

APPENDIX 3. WORKLOAD EVALUATION ANNOTATED REFERENCES (CONTINUED)

Workload and Error

1. Test Operation Procedure (TOP) 1-2-610, May 1990, Section 5.12 Test Procedure-Error Likelihood Analysis..

This document contains an excellent discussion that describes a technique for analyzing the adequacy of the design of controls and displays in relationship to human error. It correlates error likelihood rates with the degree to which the controls and displays conform to human factors design criteria. A basic premise of this approach is that error rates and workload will increase with designs that fail to properly address existing human factors design criteria. A simplistic but effective approach is to measure error probability likelihood by evaluating human factors systems design criteria, developed in a checklist format, and then compare that against the actual cockpit/system design under evaluation. The Error Likelihood Analysis section, Design Checklists and Workload Assessment techniques has been made available through the human factors web site for the small airplane directorate at the following url

http://www.faa.gov/certification/aircraft/aceMeasurementTech.pdf

In addition, the entire report can be obtained from the DTIC web site at http://handle.dtic.mil/100.2/ADA226480.

2. GAMA Publication No. 10. Part 23 Best Practices for Cockpit/Flight Deck Design

This document contains several examples of good practices to employ in designing the user interface to reduce human error. It also identifies design practices to consider in order to mitigate the consequences of errors. The document includes discussion of how to use usability testing to assist in reducing error in the design of the system and reduce the seriousness of errors made in the cockpit. This document is also available from the GAMA web site, http://www.gama.aero/pubs/itemDisplay.php?catalogID=26

3. Part 23 Human Factors Policy

This document provides several discussions regarding what to consider when performing human factors cockpit evaluations to include examining human error and the consequences of those errors. This document is available from the Small Airplane Directorate web site under human factors.

4. DOT/FAA Human Factors design guidance DOT/FAA/CT-96/1, JAN 15, 1996

This design guide discusses human error in Chapter 5. The guidance recommends assessment of complex systems that address human reliability, failure and risk assessment. It recommends that systems analysis treat humans as major components, consider human interfaces as contributing elements and treat human performance. The guidance also addresses the importance of error tolerant designs and proposes evaluation of complex systems through rapid prototyping, analytic methods or modeling.

5. Rouse, W. and Rouse, S (1983). Analysis and classification of human error. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13(4), 539-549.

Abstract: The literature on human error is reviewed with emphasis on theories of error and classification schemes. A methodology for analysis and classification of human error is then proposed which includes a general approach to classification. Identification of possible causes and factors that contribute to the occurrence of errors is also considered. An application of the methodology to checklists in the aviation domain is presented for illustrative purposes. (60 References)

APPENDIX 3. WORKLOAD EVALUATION ANNOTATED REFERENCES (CONTINUED)

6. Swain, A. and Guttman, H. (1983). Handbook of human reliability analysis with emphasis on nuclear power plant applications (NUREG/CR-1278). Washington, D.C.: Nuclear Regulatory Commission.

This handbook describes methods, models, and estimated human error probabilities that allow an analyst to make quantitative or qualitative assessments of occurrences of human errors. These examples were developed for operations involving operation of nuclear power plants. This handbook also shows how to recognize error likely-equipment design, procedures, and other human factors problems so that improvements and interventions can be considered. Although written specifically for nuclear power plants much of the information is applicable to designs of other equipment including those developed for aviation.

APPENDIX 4. RECOMMENDED TEST PLAN GUIDANCE AND DATA COLLECTION INFORMATION

1. DATA COLLECTION TECHNIQUES - Guidelines concerning the implementation of a selected number of subjective, physiological, and performance workload measurement techniques, are contained in the FAA sponsored report "Assessment of Crew Workload Measurement Methods, Techniques, and Procedures" Vol. II (Report No. WRDC-TR-89-7006).

a. Selecting Performance Measures, Defining Criteria. The criteria to evaluate the system against are set either by the performance of the reference system or by the acknowledged intended function of the system. Performance measures are developed directly from the criteria.

b. Questionnaire Development

(1) Construction/Selection of Workload Scales. Numerous workload rating scales have been developed over the years.

(2) Wording/Biasing Considerations and Capturing Key Aspects of System Workload Contributors. The structure and manner in which questions are phrased can significantly affect responses.

c. Types of Evaluations – Comparative Versus Absolute Evaluations

(1) Obtaining baseline data and making comparisons using each participant as a comparative evaluator (test-pilot evaluation)

(2) Obtaining individual performance data for a baseline and a new system

(3) Collection and Analysis of Error Data

2. FORMS. Data collection forms should be developed and provided to the FAA prior to the start of testing to demonstrate compliance to 14 CFR, part 23, § 23.1523. The applicant and the FAA should work together to ensure that data collection forms collect pertinent, meaningful data that can be used to draw conclusions about minimum crew, system workload and usability. It is generally recommended that forms be developed to query and record pilot ratings and observations, and to record in-flight observations, to include crew performance, crew errors, missed communications, and problems with checklists and flight management or flight guidance systems. When possible, the airplane should be configured to allow observers to view crew actions and hear all (i.e. internal and external) communications. It is preferable to collect data immediately after the performance of a task, as the greater the period of time between an event and the recording of data, the greater the likelihood information may be inaccurate or lost. Pilot comments are an excellent source of data; pilot comments should be encouraged and recorded throughout testing. Also, particularly during the performance of more complex tasks, pilots should be encouraged to speak their thoughts out loud as they step through the performance of a task. Members of the certification team who will serve as evaluation pilots and observers should be familiar with the data collection forms that will be used for the project prior to the start of testing. It is highly recommended that a pretest be conducted using proposed data collection forms to ensure that forms are usable and effective.

3. DEBRIEF. After the completion of each test flight, a post-flight debrief should be conducted. The test conductor should review the events of the flight and go over recorded comments with the pilot. The pilot should be given the opportunity to clarify comments and discuss salient issues with the design/system. If a post-flight questionnaire is to be given, it should be provided to the pilot prior to the post-flight debriefing.

4. DATA REDUCTION. The data collected during the course of testing will need to be reduced and summarized. Typically, statistical analysis of data is not warranted due to small sample sizes and limited testing. It is preferable to group data by topic and summarize the results in tables, indicating, for example, the number of responses for each rating (e.g. 1-Much Less Than, 2-Somewhat Less Than, 2-Equal To, 1-Somewhat More Than, and 0-Much More Than). Likewise, comments and observations should be grouped by topic and differentiated based on positive

APPENDIX 4. RECOMMENDED TEST PLAN GUIDANCE AND DATA COLLECTION INFORMATION (CONTINUED)

versus negative comments. All areas of "increased" or "unacceptably high" workload must be clearly identified and investigated. It should be noted that a finding of increased workload (over a reference system) does not necessarily constitute an unacceptable situation (certification issue), but rather indicates an area requiring further examination to ensure that workload is acceptable. The report should include a copy of all completed data collection forms and any recorded comments or observations. The applicant should provide a disposition in the report for each certification issue identified during testing. The applicant may choose to provide the test results only and permit the FAA draw the conclusions, or propose conclusions from the data collected. In the event the applicant provides conclusions in the test report, it needs to be clear that the FAA is the final authority in terms of the acceptability of the design and compliance findings.

5. LEVEL OF TESTING. The level and amount of testing should be commensurate with the level of changes and differences between the previous system (if there is one) and the new system/cockpit. In case where only a minor change has been made to a component or cockpit configuration, or where a single piece of equipment has been added to an existing approved configuration, workload analyses should be tailored to focus on the function and operation of the specific component and its interactions with other cockpit equipment. However, if there have been significant changes made to the cockpit configuration, then more extensive workload evaluations should be conducted.

6. QUESTIONNAIRE DEVELOPMENT

a. This section is not meant to be a tutorial on questionnaire construction, but rather to highlight some of the areas and issues that should be considered when developing a questionnaire for collecting data on crew workload and performance

b. A questionnaire is a subjective measurement tool for systematically eliciting perceptions or attitudes from selected personnel. Questionnaires often form an important part of the data collection effort. The major benefit of a questionnaire is that each test participant is asked the same questions in the same way, making it easier to compare results from different participants. It also forces the participants to focus on and critically evaluate the key areas and issues with a new design/system. When properly designed, the questionnaire also aids in the data tabulation and analysis. Questionnaires can be used to assess a wide variety of qualitative variables such as acceptance, ease of use, and preference. Unfortunately, some of the very characteristics that make questionnaire use so beneficial can also be detrimental.

c. Poorly researched and constructed questionnaires can cause evaluators to focus on less important (less critical) system attributes and therefore ignore or dismiss some of the key issues with the system. One of the more important areas, and also the most difficult, of questionnaire construction is wording the questions. Most authorities agree that faulty or inappropriate question wording is the greatest source of error in the use of questionnaires. Misunderstanding and misinterpretation of questions by test participants often cause errors and distortions in the resulting data. Such misunderstanding is frequently due to improper vocabulary or ambiguous phrasing.

d. Wording must be carefully selected to stimulate thought and solicit answers relevant to the system under evaluation. It is usually best if questions are worded such that there is no hint or suggestion of a "desired" answer. Questions should be neutrally stated, signifying that any answer is acceptable. For example, a question concerning the performance of a task could be stated as: Capability to insert user-defined waypoints using system "Y" was: Much Easier, Somewhat Easier, About the Same, Somewhat More Difficult, Much More Difficult than performing this task with system "X". The evaluator would be asked to circle the applicable phrase. A second method for constructing questions is to provide a mix (usually randomly distributed) of positive and negative statements and request the evaluator to rate their level of agreement or disagreement with the statement. For example, It was much easier to insert a user-defined waypoint in system "Y" than it was with system "X". The evaluator is asked to rate their level of agreement or disagreement with that statement using a scale, which may be worded: Strongly Agree, Somewhat Agree, Neither Agree or Disagree, Somewhat Disagree and Strongly Disagree.

APPENDIX 4. RECOMMENDED TEST PLAN GUIDANCE AND DATA COLLECTION INFORMATION (CONTINUED)

e. Even with the most carefully developed questionnaires, there are bound to be some areas that are missed or not adequately covered. It is, therefore, recommended that every questionnaire include a comment form, to allow the evaluator to freely express his/her opinion on system attributes. Also, another technique for collecting feedback from participating personnel is to provide a "Recommended Improvements" form. This approach may aid in identifying areas where improvements may be made, but do not constitute enough of an issue to warrant mentioning elsewhere.

7. RATING SCALES

a. Rating scales are an effective method to quantify evaluator responses. Many different types scales have been developed over the years for a variety of purposes. Typically, descriptor adjectives are used to define scale endpoints and intermediate graduations. Descriptor adjectives may characterize the quality (good to bad) of a system attribute; provide a measure of equivalency to another system (better than, about the same, worse than) or distinguish levels of adequacy or acceptability (e.g. Very Acceptable, Very Unacceptable). It is not the FAA's goal to precisely quantify crew workload or performance, but rather to ensure that the system provides for "acceptable" levels of crew workload and performance, and that no unsafe conditions exist. Consequently, rating scales can be less sophisticated. Some examples of acceptable rating scales are:

Completely Acceptable	Completely Adequate
Reasonably Acceptable	Reasonably Adequate
Borderline	Borderline
Extremely Unacceptable	Completely Inadequate
Considerably Better	Much Better
Somewhat Better	Somewhat Better
About the Same	No Better – No Worse
Somewhat Worse	Somewhat Worse
Considerably Worse	Much Worse
Much Less	Very Easy
Somewhat Less	Somewhat Easy
Same	Borderline
Somewhat More	Somewhat Difficult
Much More	Very Difficult
Very Good	Strongly Agree
Good	Largely Agree
Fair (or So-So)	Neutral
Poor	Somewhat Disagree
Very Poor	Strongly Disagree

b. A significant number of workload rating scales have been developed and used by government and industry alike. Many of these scales have been well tested and validated to be effective measure of workload, being very sensitive to slight differences in perceived workload. Any of the scales discussed in the "Assessment of crew workload measurement methods, techniques and procedures," (Volumes 1 and 2), may be used to collect workload data. Additionally, rating scale descriptors discussed previously may be used as well providing that questions are appropriately phrased for the selected scale descriptors. Figure 1 below is an example of a data collection form used to collect pilot-vehicle interface and workload data. Figures 2 and 3 provide examples of forms that may be used to collect pilot/user opinions and feedback.

FIGURE 1 – FORM TO COLLECT PILOT VEHICLE INTERFACE AND WORKLOAD DATA

Test Point No.: 1	Test Objective: Unusual Attitude Recovery Over Water - Evaluate the capability of the pilot to recognize current aircraft situation and take timely appropriate corrective action to recovery from an unusual attitude.				
Test Procedure: While over water with only water in view on the Vision 1 system, the safety pilot will put the aircraft in an unusual attitude (nose down) while the test pilot looks at his/her lap. Test pilot to recover aircraft to level flight with minimum altitude loss					
Altitude: ≥3000ft agl	**Airspeed:** As required	**Other Conditions:**			
Question		**Rating Values**			
		VG	**G**	**F**	**P**	**VP**
1. Capability to recognize relative aircraft attitude and situation at start of test:						
2. Capability to recognize, at a glance, the horizon:						
3. Performance of display for unusual attitude recovery:						
4. Workload Rating (Check appropriate box)		**Much Less**	**Less**	**Same**	**More**	**Much More**

Rating scale values are VG- Very Good, G- Good, F- Fair, P- Poor, and VP- Very Poor.

Pilot comments: _____

APPENDIX 4. RECOMMENDED TEST PLAN GUIDANCE AND DATA COLLECTION INFORMATION (CONTINUED)

FIGURE 2 – COLLECT PILOT/USER COMMENTS AND FEEDBACK

COMMENTS

Pilot's Name: _____ Date: _____ Phone: _____

General Complexity and Usability:

Option/Function Selection Procedures:

Data Entry Procedures:

System Feedback:

Display Characteristics:

Warnings, Cautions and Alerts:

Other (Specify) _____ :

APPENDIX 4. RECOMMENDED TEST PLAN GUIDANCE AND DATA COLLECTION INFORMATION (CONTINUED)

FIGURE 3 - COLLECT PILOT/USER OPINIONS AND FEEDBACK

Recommended Improvements

Pilot's Name: _____ Date: _____ Phone: _____

Design and Operation of Controls:

Option/Function Selection Procedures:

Display Characteristics:

Warnings, Cautions and Alerts:

Other (Specify) _____ :