AFRL-RH-WP-TR-2008-0054

Development of a
Supervisory Control Rating Scale

Gavan Lintern, Ph.D.
Thomas Hughes

General Dynamics
Advanced Information Systems
3200 Springfield Pike, Suite 200
Dayton OH 45431

January 2008
Interim Report for June 2006 to November 2007

Approved for public release;
Distribution is unlimited.

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
System Control Interfaces Branch
Wright-Patterson AFB OH 45433

NOTICE AND SIGNATURE PAGE

TECHNICAL REVIEW AND APPROVAL

AFRL-RH-WP-TR-2008-0054

THIS TECHNICAL REPORT HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION.

FOR THE DIRECTOR

//**signed**//	**//signed//**
Thomas R. Carretta	Daniel G. Goddard
Engineering Research Psychologist	Chief, Warfighter Interfaces Division
System Control Interfaces Branch	Human Effectiveness Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 05 Jan 2008	**2. REPORT TYPE** Interim	**3. DATES COVERED** *(From - To)* June 2006 – November 2007

4. TITLE AND SUBTITLE
Development of a Supervisory Control Rating Scale

5a. CONTRACT NUMBER
F33615-01-D-3105/D00058

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S
Gavan Lintern Ph.D.
Thomas Hughes

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
17

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
General Dynamics
Advanced Information Systems
3200 Springfield Pike, Suite 200
Dayton OH 45431

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Material Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
System Control Interfaces Branch
Wright-Patterson AFB OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RHCI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RH-WP-TR-2008-0054

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; Distribution is unlimited.

13. SUPPLEMENTARY NOTES

88 ABW PA Cleared 04/30/08, WPAFB-08-3139.

14. ABSTRACT
The objective was to develop a supervisory control rating scale to evaluate human interaction and capabilities associated with automation. The product was to be a standardized rating scale analogous to the Cooper-Harper Rating Scale that was developed to assess aircraft handling response. The use intended for the scale developed under the current project is to evaluate supervisory control across situations and human-system interface concepts in a manner that reflects supervisor's workload, situational awareness, and complacency. The report summarizes a review/analysis of related literature and the initial scale development process. Additional studies are required to examine the psychometric properties of the scales including sensitivity, reliability (internal consistency, inter-rater, test-retest), and construct validity.

15. SUBJECT TERMS
Supervisory control, unmanned air systems, situational awareness, workload

16. SECURITY CLASSIFICATION OF: **Unclassified**			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Thomas R. Carretta
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	33	**19b. TELEPONE NUMBER (*Include area code)***

THIS PAGE INTENTIONALLY BLANK

CONTENTS

FIGURES

TABLES

PREFACE

This report describes activities performed under Contract F33615-01-D-3105/D00058, Task 007 in support of the Air Force Research Laboratory Warfighter Interface Division, System Control Interfaces Branch (AFRL/RHCI) Interfaces for Small Unmanned Systems, Work Unit 71840917.

THIS PAGE INTENTIONALLY LEFT BLANK

Introduction

The stated project objective was to develop a supervisory control rating scale to evaluate human interaction and capabilities associated with automation. The product was to be a standardized rating scale analogous to the Cooper-Harper Rating Scale that was developed to assess aircraft handling response. The use intended for the scale developed under the current project is to evaluate supervisory control across situations and human-system interface concepts in a manner that reflects supervisor's workload, situational awareness, and complacency. There is consensus in the project team that the application area most suitable for illustrating the use of the new rating scale is supervisory control of unmanned air vehicles (UAVs).

As background to this development effort, we have reviewed one of the original papers that describes the Cooper-Harper Rating Scale (Harper & Cooper, 1986) and some more recent work out of the Cranfield University in the United Kingdom that, although based on the Cooper-Harper scale, was aimed at developing a more valid and reliable instrument. We summarize those papers and their implications for this project below.

Without losing sight of the fact that the primary goal is to develop a rating scale to evaluate human interaction and capabilities associated with automation and to assess a supervisor's workload, we thought it important to develop an understanding of the background material that is found in the traditional areas of supervisory control and automation. We reviewed a number of papers in those areas and summarize the insights gleaned from them below in advance of summarizing developments of the Cooper-Harper Rating Scale. Finally in this report, we describe two forms of rating scales that we believe satisfy project requirements.

Supervisory Control

Sheridan (1988) identified supervisory control as a scheme by which an automatic sub-system uses sensed information about the state of an ongoing physical process in conjunction with information programmed into it by a human supervisor to direct actions on that process (See Figure 1). The human supervisor works through the computer to effect what needs to be done in the physical world. The computer is a mediator, updating the supervisor as it controls the physical process. The concept of supervisory control has been applied primarily to vehicular, process and robotic control. Its application to control of multiple UAVs would seem to be a reasonable extension.

Levels of Automation Support

In supervisory control, the human operator relinquishes responsibility for direct control to an automatic sub-system and takes on the role of monitor and goal-constraint setter. From his perspective that people and machines are complementary, Sheridan (1988, 1997) emphasizes a hierarchy for categorizing human activities in supervising physical processes. Sheridan detailed ten functional levels (see Table 1)in which each level specifies a balance between roles for computerized automation and the human supervisor, ranging from complete automation in which the human has no role to a complete lack of automated assistance in which the human does everything.

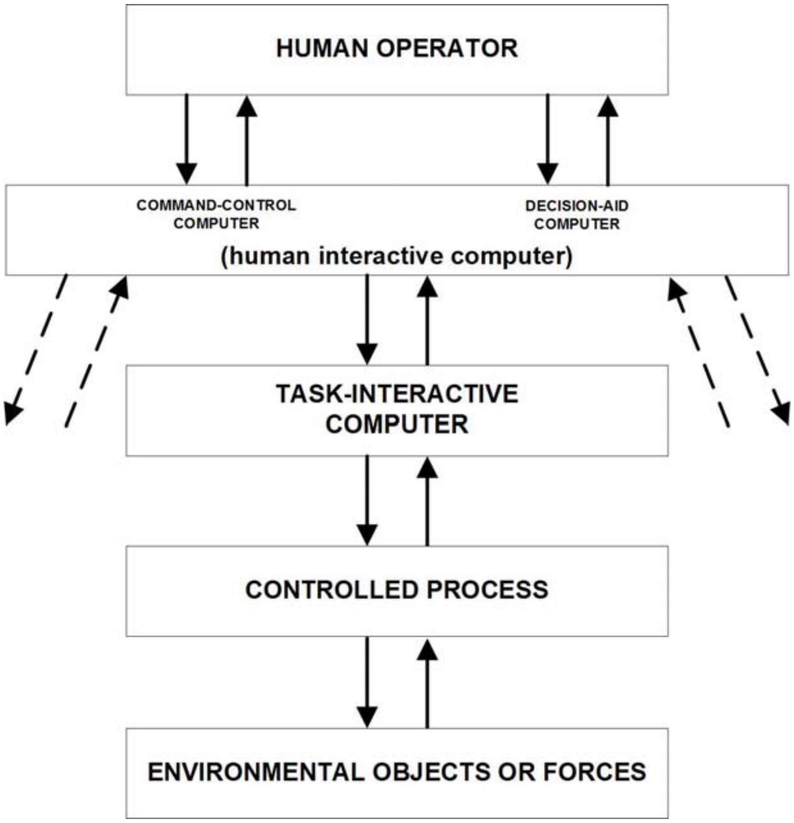
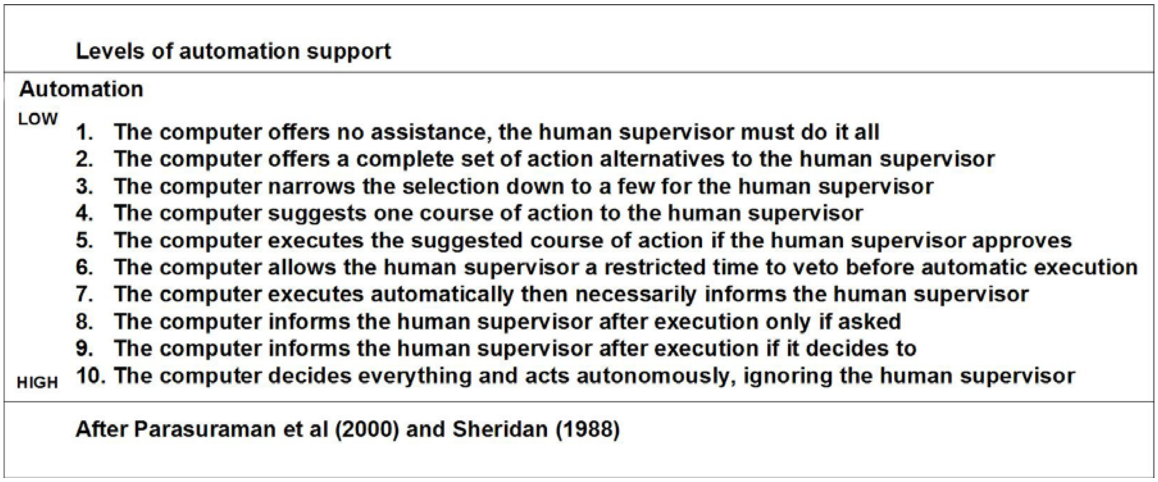

HUMAN OPERATOR
COMMAND-CONTROL COMPUTER
DECISION-AID COMPUTER
(human interactive computer)
TASK-INTERACTIVE COMPUTER
CONTROLLED PROCESS
ENVIRONMENTAL OBJECTS OR FORCES

Figure 1. A supervisory control system (from Sheridan, 1988)

Table 1. Levels of Automation Support



Levels of automation support

Automation

LOW

1. The computer offers no assistance, the human supervisor must do it all
2. The computer offers a complete set of action alternatives to the human supervisor
3. The computer narrows the selection down to a few for the human supervisor
4. The computer suggests one course of action to the human supervisor
5. The computer executes the suggested course of action if the human supervisor approves
6. The computer allows the human supervisor a restricted time to veto before automatic execution
7. The computer executes automatically then necessarily informs the human supervisor
8. The computer informs the human supervisor after execution only if asked
9. The computer informs the human supervisor after execution if it decides to
10. The computer decides everything and acts autonomously, ignoring the human supervisor

HIGH

After Parasuraman et al (2000) and Sheridan (1988)

One of the issues generally ignored in relation to the use of this classification is that it is descriptive, but not prescriptive. While any supervisory control system can be classified as belonging to one of the ten levels, the classification scheme does not indicate what balance of human versus automation is desirable for any given situation. Although it is a useful pedagogical device, it does not guide design.

Parasuraman, Sheridan, and Wickens (2000) extended this classification by crossing it with the four information-processing categories of information-acquisition, information-analysis, decision-selection, and action-implementation based on a simple four-stage model of human information processing (See Figure 2). Parasuraman et al (2000) suggested their model can be used prescriptively to guide the design of automation within each of the four information-processing categories, but their discussion of design principles is vague and is not tied to their descriptive classification. They hedge this issue by acknowledging that there is no simple answer to the question of what level of automation should be applied within each category. We judge this effort as incomplete and as yet, now seven years later, the Parasuraman strategy does not appear to have developed into an active research program.

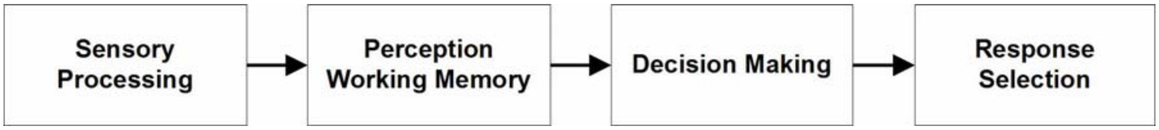


Figure 2. A simple four-stage model of human information processing (from Parasuraman et al, 2000)

The Substitution Myth of Function Allocation

Dekker and Woods (2002) took Parasuraman et al (2000) to task not because of the imprecision of their guidelines but because of their putative reliance on a function allocation paradigm, which Dekker and Woods regard as misguided. They argued that function allocation is driven by a substitution myth that people and computers have fixed strengths and weaknesses and that the task of a designer is to capitalize on the strengths while eliminating or compensating for the weaknesses. Dekker and Woods correctly point out that allocation of a particular function to automation has consequences beyond absorption of that function into the system. New functional demands are created for the other partner in the human–machine equation and these can radically modify the work demands.

Dekker and Woods (2002) argued that humans and automation should be viewed as team players and so their proposal for successful automation relates to how to support the coordination between people and automation, not how functions are distributed between them. Dekker and Woods offer some valuable principles for supporting the use of (coordination between people and automation), such as:

- Highlight changes and events in ways that the current generation of state oriented displays do not

- Use historical information to help human operators of dynamic systems anticipate what to expect and where to look next
- Use pattern- or form-based representations to convert arduous mental tasks into straightforward perceptual ones

Nevertheless, they do not address the issue of how to specify the functionality of the automation.

We were at first puzzled by the Dekker and Woods (2002) critique. Parasuraman, et al (2000) sought to develop a strategy for functional assignment while Dekker and Woods (2002) focused their comments on issues of collaboration between humans and automation and on the visibility of automated processes. These are separate issues. Only in their last paragraph do Dekker and Woods (2002) clarify their position unambiguously. They argued that functional assignment is not a relevant issue. It is not that Parasuraman, et al (2000) have pursued a flawed strategy, but have gone astray by even seeking to address the issue of functional assignment.

Dekker and Woods (2002) position is mystifying. It remains unclear how one can proceed with the design of anything without thinking about functionality. It is always possible to accept the existing or specified functionality and to then work on coordination and visibility, but that only transfers the functional assignment problem to someone else[1]. Unfortunately, Dekker and Woods failed to address this issue and offered no substantive argument for their claim that "system developers should abandon the traditional 'who does what' question of function allocation" (p 243).

The Techno-Centric Approach to Automation

The development of automation (and more generally, of human systems interaction) has been plagued by a techno-centric worldview as enunciated by Birmingham and Taylor (1954) in their observation that "man is best when doing least". This sort of view has encouraged the predominant engineering strategy of automating what can be conveniently automated and leaving the rest to the human supervisor-controller[2]. It is difficult to imagine how this perspective has gained currency among designers who, being human themselves, must realize that they do not function well if they are constrained to doing as little as possible. We suggest, in contrast, that man is best when mindful (Weick & Sutcliffe, 2001) and engaged with ongoing work processes in a manner that takes account of both local and global constraints (Lintern, 2007).

The groundbreaking research of Sarter and Woods (1992, 1994) has revealed the poverty of the techno-centric worldview as applied to the development of automation. Their work showed that an unthinking approach to automation in commercial airline cockpits reduced pilot workload in flight modes that were normally benign (where

[1] We discount the possibility that Dekker and Woods (2002) have inadvertently overstated their case from the fact that David Woods submitted, to a recent symposium panel, the proposal that function allocation is a bad idea that will not go away.

[2] This appears to have been the dominant strategy for designs of the early glass cockpits.

assistance was least needed) and increased it in flight modes that were normally challenging (where assistance was most needed).

Parasuraman et al (2000) allow that the techno-centric strategy (automate everything that one can) is a viable strategy when efficiency or cost reduction are driving forces. However, we disagree. It imposes a huge risk that efficiency or cost reductions will be undermined by *clumsy automation* of a type that creates opportunities for new kinds of human error and new paths to system breakdown. Woods and Patterson (2000), for example, note that although automation is often justified on the grounds that it helps offload work from harried practitioners, it in fact creates additional tasks, forces new cognitive strategies, and demands more knowledge and more communication. More often than not, these new demands are imposed at the very times that practitioners are most in need of true assistance (Sarter & Woods, 1992, 1994; Sarter, Woods & Billings, 1997).

At first glance, it may seem that the emergence of *clumsy automation* is an unfortunate and perverse coincidence. We suggest, however, that it is inevitable where the design of automation is guided by a techno-centric worldview. It is during the demanding times that operators are most heavily involved in activities that are difficult to automate, a point that those who wish to automate everything possible have not yet noticed.

The strategy of making automation more human-centered as proposed by Dekker and Woods (2002) is likely to ameliorate the demands at these difficult times. Their approach does not, however, encapsulate the systems view that is necessary to establish a comprehensive solution. We propose that a comprehensive, systems-oriented solution can be found in a work-centered analysis that identifies the functional demands and then establishes a suitable functional and organizational work structure in which technology is designed to support human cognitive work.

A Work-Centered Approach to Automation

Sheridan (1988) stated that people and machines are complementary, a view that Parasuraman et al (2000) maintain at least implicitly. In contrast, we suggest that this view serves to sustain the techno-centric worldview and that automation should rather be viewed as a tool or a functional support. From this human-centric perspective, automation takes a subsidiary role. The proposal by Christoffersen and Woods (2002) that humans and automation should be viewed as team players is posed as a break from the techno-centric tradition but still accords equal status to the two types of agents, thereby leaving open to the troubling issue of who should be in charge (Inagaki, 2003).

We bring a contrasting work-centered perspective to this problem: work is the responsibility of human agents where technology has the specific purpose of supporting those human agents in their work. Thus, all forms of automation are tools for the support of human work. We would not, for example, consider an automobile driver and the automatic transmission or the cruise control as members of a team. We rather would consider this system to have a human driver who has access to supports, tools, or assistive automation. From this perspective, the human is always the responsible party and the conundrum posed by Inagaki (2003) has no relevance.

This work-centered perspective forces a new strategy of deciding on the functionality of automation. It demands a much deeper analysis of the nature of the work, its functional purpose, its organization and its processes. In short, it demands a systems view of human work.

From this perspective, we find the development offered by Parasuraman et al (2000) to be well motivated. We regard function allocation as an important issue for the design of Human-Systems interaction but, in concert with Parasuraman et al (2000), we view previous strategies of function allocation as limited. We are not, however, persuaded that Parasuraman et al. have proceeded in a productive direction although, in contrast to Decker and Woods (2002), we would prefer to suspend judgment of their approach pending further developmental work in this area. While the outright rejection by Decker and Woods of function allocation strikes us as bizarre, we otherwise resonate with what they say about collaboration and the visibility of automated processes. Hollnagel and Woods (2005) and Woods and Hollnagel (2006) offer more detail on these ideas.

To summarize our work-centered perspective as it relates to this problem, supervisory control is about supervisors coordinating with other people and also coordinating the support functions of technological sub-systems. Supervisory management is possibly a more appropriate term. Note that in contrast to Christoffersen and Woods (2002), we do not think of human supervisors coordinating with technological subsystems (i.e., we do not think of humans and technology as team players), but rather think of human supervisors as using their technological subsystems (i.e., coordinating those supporting functions into their work activities).

Further, we seek to expand the focus to take in forms of cognitive support beyond automation. While automation will remain important, important forms of cognitive support can be generated by appropriate displays of information, well integrated communication tools, and support structures for organizing workflow. A comprehensive approach to supervisory control needs to take all potentially useful forms of cognitive support into account.

A rating scale for supervisory control would ideally assess the effectiveness of the coordination between the human participants in the system and how well their cognitive support tools satisfy their needs. One dimension of this assessment would be devoted to evaluating whether the subsystems and cognitive support tools have the desired functionality. Another dimension of the assessment would be devoted to evaluating conformance to principles outlined by Woods and Hollnagel (2006), for example, how visible the processes embedded in cognitive support tools are to the user, how well historical information is displayed, and whether the use of pattern of form-based representations is effective in converting arduous mental tasks into straightforward perceptual ones.

The Cooper-Harper Scale and its Extensions

Motivation

One of the primary motivations for development of the Cooper-Harper scale was found in the observation that analysis of the open-loop aircraft response is not a good guide to its human-in-the-loop handling response (Harper & Cooper, 1986). Although an open-loop analysis might suggest that dynamic response is well behaved, the human-in-the-loop handling response could be unsatisfactory if pilot-induced oscillations are prevalent during closed-loop control.

The potential for instability will vary with the nature of the task. A tight-feedback, high-gain task such as instrument-referenced glide-path tracking is more likely to introduce instability than, for example, a low-gain instrument-referenced descent even though the demanded control actions and aircraft performance are similar. Harper and Cooper (1986) emphasized that a handling quality rating is not assigned to the aircraft itself but to the aircraft as it responds within a specific flight maneuver.

Thus, to rate aircraft response effectively, one must understand in depth how the aircraft is going to be used in the full range of flight regimes. Harper and Cooper (1986) argue that pilots close different feedback loops during different tasks and even during different portions of tasks. Handling quality can therefore vary considerably for different phases of a task. It is also likely to vary with use of different control strategies even for the same task phase as, for example, use in closed-loop tracking of a high-gain control strategy to follow perturbations closely versus a low-gain strategy to dampen human-induced oscillations. For our project, this suggests that we need to identify both the essential dynamic properties of the systems to be rated, the work situations and work problems that must be faced, and the strategies that might be useful within that work domain.

Aeronautical engineers presumably know enough about human-in-the-loop characteristics to avoid the worst forms of instability but the fact that rigorous and extensive flight testing is undertaken attests to pervasive acknowledgment that systematic design cannot guarantee a stable aircraft. Similarly, with supervisory control, there are a host of important design principles that guide the convergence towards a usable system, but human-in-the-loop evaluation remains as the only trustworthy method of ensuring usability. In that respect, the assessment challenge for the two different systems is similar. Given that the Cooper-Harper scale has found widespread use in evaluation of aircraft controllability, the interest in development of a supervisory control rating scale similar to the Cooper-Harper scale seems well motivated.

Cooper-Harper Scale Description

The Cooper-Harper scale is reproduced in Figure 3. It was developed for use by test pilots and flight test engineers in evaluating the handling qualities of aircraft during flight tests. The scale values range from 1 to 10, with 1 indicating the best handling characteristics and 10 the worst.

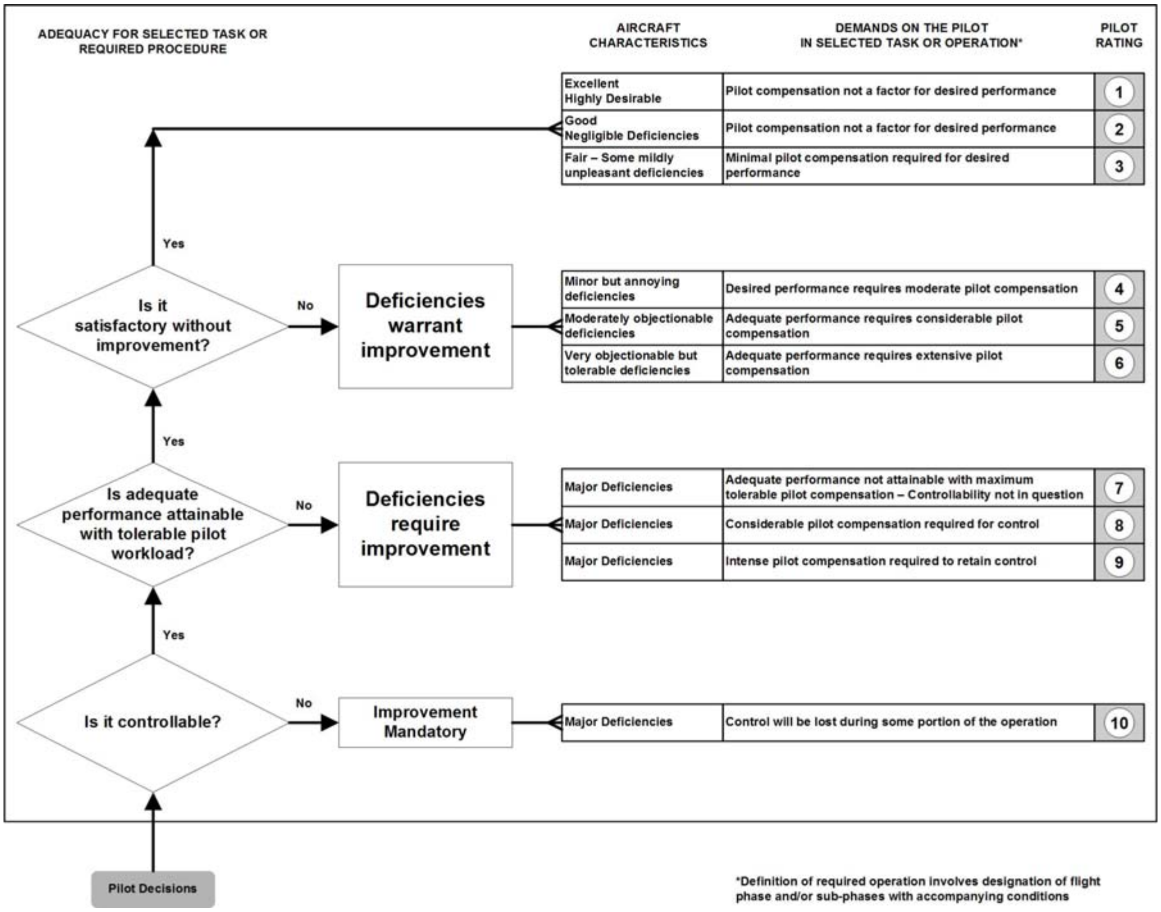

ADEQUACY FOR SELECTED TASK OR REQUIRED PROCEDURE
AIRCRAFT CHARACTERISTICS
DEMANDS ON THE PILOT IN SELECTED TASK OR OPERATION*
PILOT RATING
Excellent Highly Desirable
Pilot compensation not a factor for desired performance
1
Good Negligible Deficiencies
Pilot compensation not a factor for desired performance
2
Fair – Some mildly unpleasant deficiencies
Minimal pilot compensation required for desired performance
3
Yes
Is it satisfactory without improvement?
No
Deficiencies warrant improvement
Minor but annoying deficiencies
Desired performance requires moderate pilot compensation
4
Moderately objectionable deficiencies
Adequate performance requires considerable pilot compensation
5
Very objectionable but tolerable deficiencies
Adequate performance requires extensive pilot compensation
6
Yes
Is adequate performance attainable with tolerable pilot workload?
No
Deficiencies require improvement
Major Deficiencies
Adequate performance not attainable with maximum tolerable pilot compensation – Controllability not in question
7
Major Deficiencies
Considerable pilot compensation required for control
8
Major Deficiencies
Intense pilot compensation required to retain control
9
Yes
Is it controllable?
No
Improvement Mandatory
Major Deficiencies
Control will be lost during some portion of the operation
10
Pilot Decisions
*Definition of required operation involves designation of flight phase and/or sub-phases with accompanying conditions

Figure 3. The Cooper-Harper handling qualities rating scale (Harper & Cooper, 1986).

Although the scale itself has only one scaling dimension, the rating format has an unusual decision-tree structure that encourages raters to incorporate multi-dimensional assessments into their ratings. The decision tree format requires that the rater make a series of up to three binary decisions about the handling qualities of the aircraft. These binary decisions classify the aircraft as:

- Controllable (Yes or No)
- Handling is adequate within tolerable pilot workload (Yes or No)
- Handling is satisfactory without improvement (Yes or No)

A negative response at the first decision point is based on a judgment that the aircraft is uncontrollable within the tested flight regime (correction is mandatory). A negative response at the second decision point is based on a judgment that adequate performance is not attainable with tolerable pilot workload (improvement is necessary). A negative response at the third decision point is based on a judgment that the aircraft has no deficiencies or some mild deficiencies (improvement may be warranted). Negative decisions at the second and third decision points and an affirmative decision at the third decision point lead to finer discriminations of the degree of pilot compensation required to maintain controllability. Given the well-established difficulty of subjectively discriminating more than seven items (Miller, 1956); ratings made on this two-tiered, nested scale are presumably more reliable than they would be on a straight 10-point scale.

The Cranfield Development

Within the past decade, the Department of Human Factors at Cranfield University in the United Kingdom has undertaken the development of another scale to evaluate aircraft handling qualities (Harris, Gautrey, Payne & Bailey, 2000; Payne & Harris, 2000). These authors have expressed concern that the Cooper–Harper scale does not identify the control dimensions on which the aircraft is being assessed. They note that the Cooper–Harper ratings reflect different but unspecified dimensions of an aircraft's handling performance depending on the nature of the maneuver being undertaken. Payne and Harris (2000) argue that because these different control dimensions are not identified explicitly in the ratings, the scale is not diagnostic. Indeed, Harper and Cooper (1986) emphasize that their scale must be used in conjunction with other notes made by the pilot, but Payne and Harris (2000) apparently find that approach too informal and sought to develop a scale that would identify the dimensions being assessed.

In developing their scale, Payne and Harris (2000) identified the flight control dimensions of pitch, roll, yaw, trim, and speed as relevant to flight performance and demonstrated by multivariate analysis that the criticality of these dimensions differed across different phases of flight. The second dimension of their scale required a criticality rating for each of these control dimensions for the phase of flight being assessed.

We should note however that the Cooper–Harper scale was developed to assess system response in a high intensity and unforgiving environment. It would be unrealistic to expect that test pilots engaged in evaluating the response of a new aircraft through demanding flight regimes could fill out anything more complex than a one-dimensional scale. Nevertheless, it would be useful to develop a scale that can be diagnostic and the Cranfield work motivated us to develop a diagnostic capability although, as discussed later in this report, we concluded that a criticality index did not translate well from assessment of flight control to assessment of supervisory control.

It is also noteworthy that Payne and Harris (2000) found reason to question the reliability of the Cooper–Harper Scale. They reference a study by Wilson and Riley (1989) that showed poor inter-rater (pilot) reliability and also a study by Field (1995) that revealed poor within-rater (pilot) reliability. Field's observation that the same pilot could vary the Cooper–Harper rating for the same aircraft configuration by several points (e.g., from 3 (satisfactory without improvement, but some minor deficiencies) to 7 (deficiencies require improvement—adequate performance not attainable with maximum pilot compensation), is particularly troubling for a scale such as the one designed by Harper and Cooper (1986). Reliability is the most fundamental property of measurement and its absence calls into question the value of the scale.

Payne and Harris (2000) also argued that the Cooper–Harper scale lacks sensitivity, although their argument on this score is not well substantiated. Essentially, they offer the general claim that multidimensional scales are more sensitive than one-dimensional scales, but do not offer evidence that the Cooper–Harper scale itself is insensitive. Nevertheless, this is an important issue. An insensitive scale will not distinguish effectively between system responses that should be distinguished. They argue that the NASA Task Load Index (Hart & Staveland, 1988) and the subjective workload assessment technique (Reid & Nygren, 1988) are both more sensitive than the Cooper–Harper scale.

Finally, Payne and Harris (2000) point out that there has been little effort to establish the validity of the Cooper–Harper scale. They question the validity of this scale because the ratings are often not consistent with the supporting notes made by the test pilots. Again referencing Wilson and Riley (1989), they observed that pilot opinions as expressed on the comment cards occasionally suggested that the aircraft handling response was unacceptable while the Cooper–Harper ratings suggested otherwise.

Given the issues identified above, we sought to develop a strategy that would be simple to use (although non-diagnostic) in a first assessment, but that could be extended into a diagnostic instrument when that was needed. Validity, reliability, and sensitivity remain as concerns, but the extensive data collection effort and statistical analysis required to establish them were beyond the scope of this project. In the work described below, we took the first steps in the development of the scale, that is, identification of the key assessment dimensions and construction of a scale format.

Rating Scale Issues

The objective of this project was development of a standardized rating scale that could be used to evaluate supervisory control across situations and human-system interface concepts. The specified requirement was to develop a set of behavioral anchors that would reflect the supervisor's ability to:

- acquire the information needed to achieve and maintain adequate levels of situational awareness,
- process and analyze the information in a timely manner pursuant to making action decisions,
- make an informed decision, and
- take appropriate actions.

We understood this to mean that the rating scale should assess how effective the supervisory controller is in relation to using a technological system for its specific mission.

The Role for an Assessment Instrument

The assessment instrument under development will be used to rate the quality of a technical system rather than the competence or skill of its users. In that role, the instrument might be employed as a usability evaluation tool or alternatively, as a diagnostic tool. The Cooper-Harper Rating Scale (Harper & Cooper, 1986), is a usability evaluation tool that distinguishes at a global level between aircraft with desirable handling characteristics and those with undesirable handling characteristics. The development of the Cranfield Aircraft Handling Qualities Rating Scale (an extension of the Cooper-Harper Rating Scale) was justified on the basis of the need for a diagnostic tool (Harris et al, 2000).

In contrast to the Cooper-Harper Rating Scale, the Cranfield Aircraft Handling Qualities Rating Scale has two scales, both of which are multidimensional. One is a 10-point scale for rating aircraft dynamic response on the five control dimensions of pitch, roll, yaw, trim, and speed. The other is a 5-point scale (criticality index) for rating how critical each of these control dimensions is for the particular flight maneuver being undertaken. The criticality index is calibrated separately for each flight maneuver so that the Cranfield instrument assesses not only the aircraft handling qualities in relation to the different control dimensions, but also the interaction between handling qualities and the demands of the flight task.

The Cranfield Scale suggests a form that would be useful in our selected domain; supervisory control of UAVs. Although the rating item content will be quite different, the strategy of assessing usability in relation to fundamental dimensions of the control task[3] on one scale and assessing the criticality of those dimensions for different phases of the work is one that seems to accommodate the basic requirements for assessment of a supervisory control system.

Formal Scale Properties

A usability scale places fewer demands on the formal properties of an assessment scale than does a diagnostic scale. While both demand reliability (test-retest and/or between-subjects), sensitivity (discriminates between systems that are more or less usable), and construct and content validity (instrument actually measures what it appears to measure), the construct and content validity demands for a diagnostic scale must effectively measures multiple dimensions of the target construct. Content validity is related to face validity, although the latter refers not to what the test actually measures but to what it appears to measure. Face validity is derived from intuitive judgment but does not attest the soundness of an instrument. The establishment of construct and content validity requires statistical comparison with other validated measures.

Dimensions of Supervisory Control

The operational definition for supervisory control hypothesized that workload, situational awareness, and complacency should be considered as related concepts. In this section some of the foundational literature on workload and situational awareness is reviewed. The literature on complacency is less definitive and we summarized a paper by Moray and Inagaki (2000) who argued that there is little evidence for complacent behavior. A paper by Cummings, Meyers, and Scott (2006) that described development of a Cooper-Harper style rating scale for display evaluation also was reviewed. Although

[3] The term “Control Task”, taken from Vicente (1999) describes what must be accomplished and the cognitive resources required. However, under the assumption that a socio-technical system permits goals to be satisfied in different ways, it does not describe an action sequence.

we concluded the work described by Cummings et al had no apparent direct relevance to this project, a summary is provided in Appendix A.

Workload

Hart and Staveland (1988) define workload as "a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance." [p 140] They further note that it is not uniquely defined by the objective task demands, that it reflects multiple attributes that may have different relevance for different individuals, and that it is an implicit combination of factors. Workload measures are classified as one of three types:

- Performance-based: performance on a secondary task is used to estimate the spare capacity available in performance of a primary task
- Subjective: rating scales are used to identify the source of excessive workload
- Physiological/biochemical: these measures are based on the assumption that level of arousal varies as a function of workload (e.g., pupillary response)

This project has focused on subjective of workload assessment. The two most frequently cited instruments in this category are the Subjective Workload Assessment Technique (SWAT; Reid & Nygren, 1988) and the NASA Task Load Index (NASA TLX; Hart & Staveland, 1988). The SWAT rates workload on three scale factors (time load, mental effort, and psychological stress) at three levels (low, medium, high). This scale is time-consuming to administer. but is said to be reliable, diagnostic (presumably because it assesses workload on three different scale factors), and sensitive although of undetermined validity (European Organization for the Safety of Air Navigation, 2003). The NASA TLX rates workload on six scale factors (mental, physical, temporal, effort, performance, and frustration) followed by a paired comparison weighting process to determine the importance of each factor for the task in question. This scale requires just a short time to administer and is said to be reliable, diagnostic (presumably because it assesses workload on six different scale factors), and sensitive. Further, the NASA TLX has been extensively validated (European Organization for the Safety of Air Navigation, 2003).

Situation Awareness

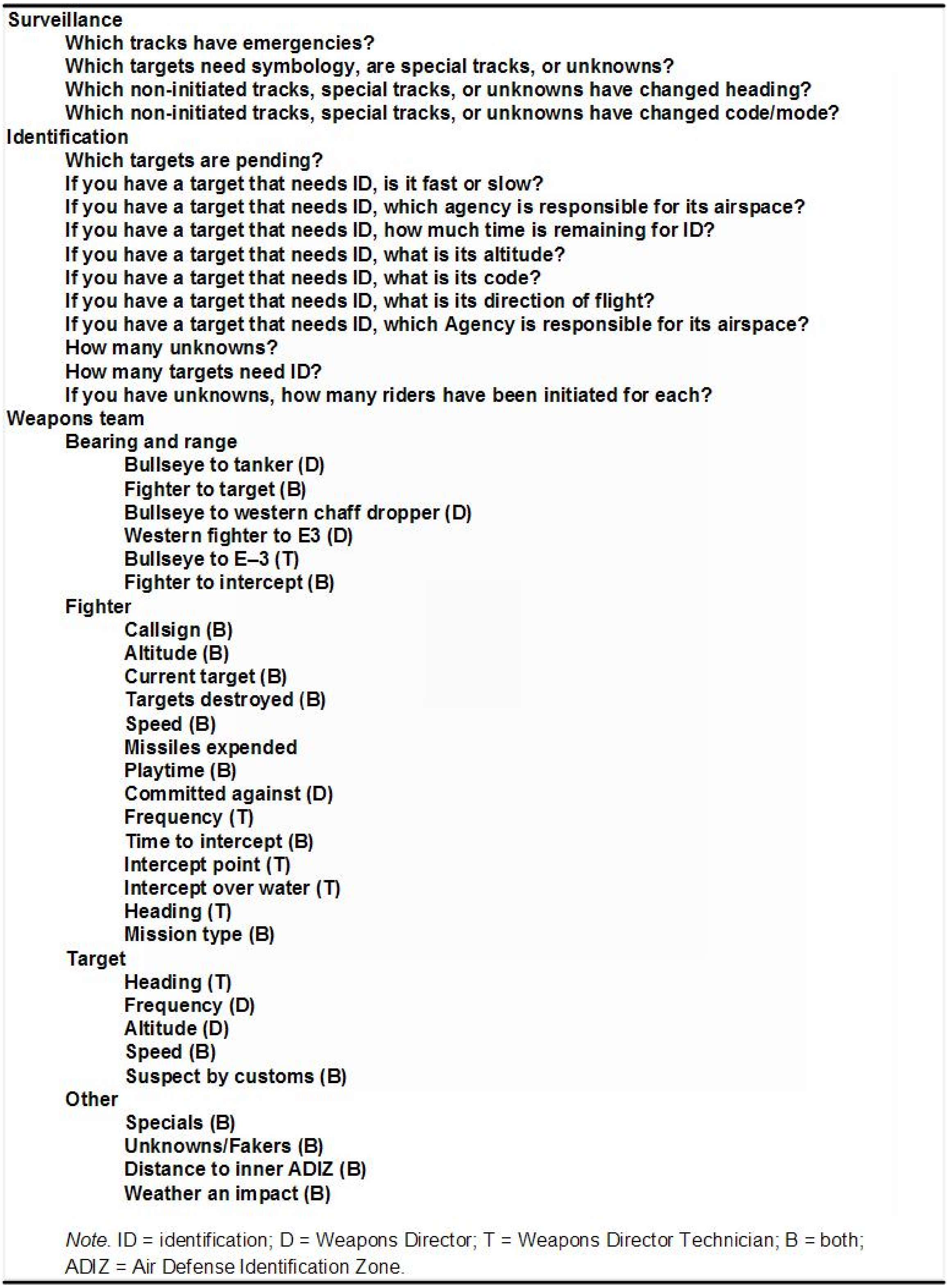
Endsley (1988) defines situation awareness (SA) as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" and has developed the Situation Awareness Global Assessment Technique (SAGAT) to measure SA during dynamic simulations. When SAGAT is used, the simulation is halted at random intervals and the controller is probed with SA-related questions. The questions probe current perceptions of the environment, comprehension of the situation, and their projections of the future. Answers typically are compared to actual states of the environment to evaluate the accuracy of a controller's situation awareness.

Two characteristics of this scale should be noted. The first is that the questions are generated from a SA analysis for a particular domain. Thus, the SAGAT must be customized for each separate domain. In that respect, it is unlike scales such as the NASA TLX workload scale that can be applied without modification to any domain. On the positive side, there is considerable guidance (Endsley, 1988, 1995; Jones & Endsley, 2004) on the type of items that should be generated. In particular, the definition of situation awareness indicates that items should assess perception of the elements in the current environment, comprehension of their meaning, and projection of their future status.

The second noteworthy feature of the SAGAT, in contrast to scales such as the NASA TLX, Cooper-Harper, and Cranfield scales, is that responses are evaluated in relation to objective states. Most scales require the controller to judge the usability of the system whereas the SAGAT, in contrast, requires the controller to assess an objective state. In the case of the SAGAT, it is the investigator rather than the controller who infers the usability of the system based on the accuracy of the probe responses.

One concern with the SAGAT is the need to halt the simulation to administer the items. This can be disruptive and interfere with the progress of the simulated mission. In addition, our scale might need to be used with a real system in which it would not be possible to halt the dynamic process. Jones and Endsley (2004) addressed this concern by comparing real-time and SAGAT probes in an air traffic control scenario. See Tables 2 and 3.

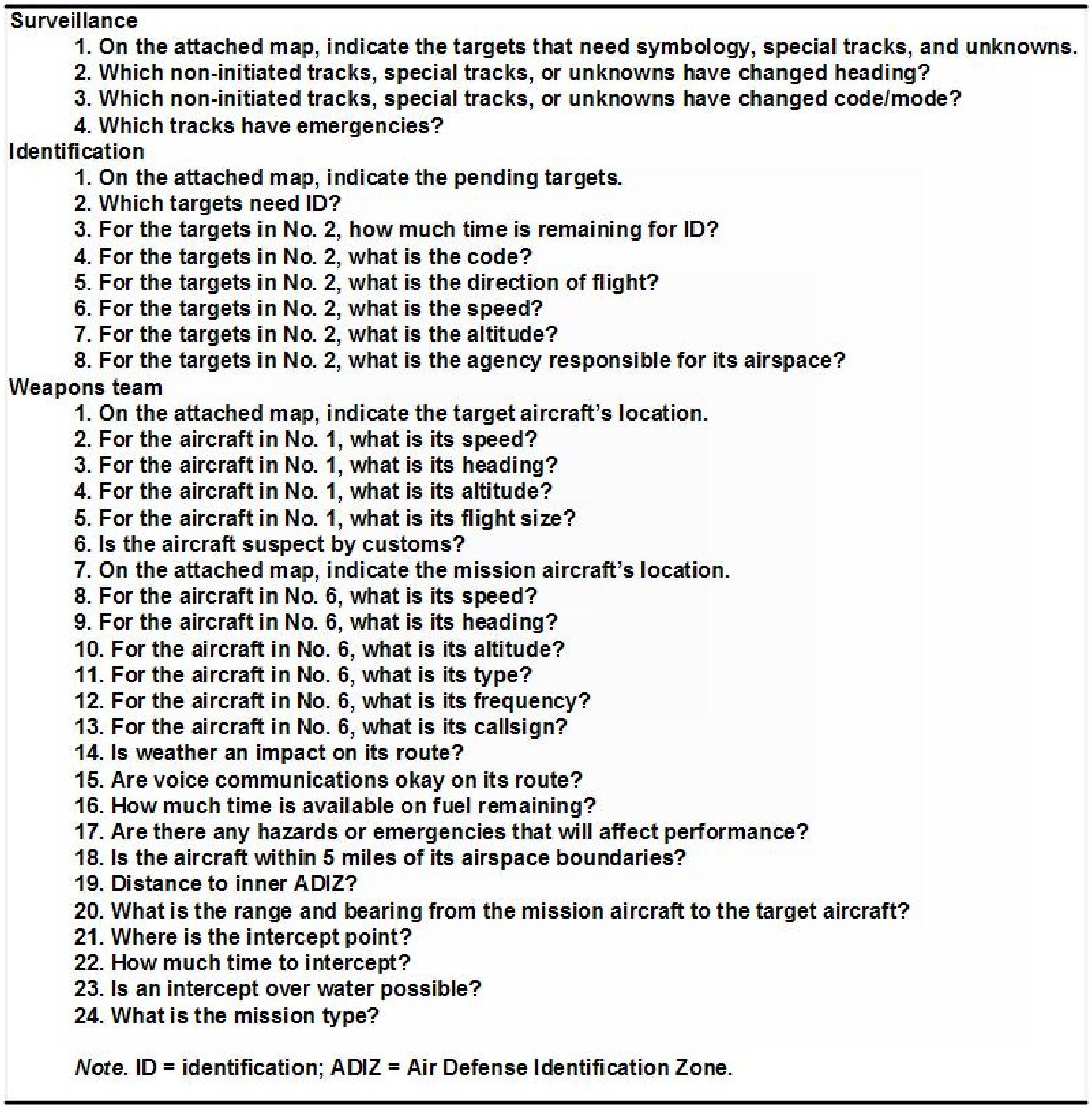
Table 2. Situation Awareness Real-Time Probes for Air Traffic Control



- Surveillance
  - Which tracks have emergencies?
  - Which targets need symbology, are special tracks, or unknowns?
  - Which non-initiated tracks, special tracks, or unknowns have changed heading?
  - Which non-initiated tracks, special tracks, or unknowns have changed code/mode?
- Identification
  - Which targets are pending?
  - If you have a target that needs ID, is it fast or slow?
  - If you have a target that needs ID, which agency is responsible for its airspace?
  - If you have a target that needs ID, how much time is remaining for ID?
  - If you have a target that needs ID, what is its altitude?
  - If you have a target that needs ID, what is its code?
  - If you have a target that needs ID, what is its direction of flight?
  - If you have a target that needs ID, which Agency is responsible for its airspace?
  - How many unknowns?
  - How many targets need ID?
  - If you have unknowns, how many riders have been initiated for each?
- Weapons team
  - Bearing and range
    - Bullseye to tanker (D)
    - Fighter to target (B)
    - Bullseye to western chaff dropper (D)
    - Western fighter to E3 (D)
    - Bullseye to E–3 (T)
    - Fighter to intercept (B)
  - Fighter
    - Callsign (B)
    - Altitude (B)
    - Current target (B)
    - Targets destroyed (B)
    - Speed (B)
    - Missiles expended
    - Playtime (B)
    - Committed against (D)
    - Frequency (T)
    - Time to intercept (B)
    - Intercept point (T)
    - Intercept over water (T)
    - Heading (T)
    - Mission type (B)
  - Target
    - Heading (T)
    - Frequency (D)
    - Altitude (D)
    - Speed (B)
    - Suspect by customs (B)
  - Other
    - Specials (B)
    - Unknowns/Fakers (B)
    - Distance to inner ADIZ (B)
    - Weather an impact (B)

Note. ID = identification; D = Weapons Director; T = Weapons Director Technician; B = both; ADIZ = Air Defense Identification Zone.

Table 3. Situation Awareness Global Assessment Technique (SAGAT) Queries for Air Traffic Control



Surveillance

1. On the attached map, indicate the targets that need symbology, special tracks, and unknowns.
2. Which non-initiated tracks, special tracks, or unknowns have changed heading?
3. Which non-initiated tracks, special tracks, or unknowns have changed code/mode?
4. Which tracks have emergencies?

Identification

1. On the attached map, indicate the pending targets.
2. Which targets need ID?
3. For the targets in No. 2, how much time is remaining for ID?
4. For the targets in No. 2, what is the code?
5. For the targets in No. 2, what is the direction of flight?
6. For the targets in No. 2, what is the speed?
7. For the targets in No. 2, what is the altitude?
8. For the targets in No. 2, what is the agency responsible for its airspace?

Weapons team

1. On the attached map, indicate the target aircraft's location.
2. For the aircraft in No. 1, what is its speed?
3. For the aircraft in No. 1, what is its heading?
4. For the aircraft in No. 1, what is its altitude?
5. For the aircraft in No. 1, what is its flight size?
6. Is the aircraft suspect by customs?
7. On the attached map, indicate the mission aircraft's location.
8. For the aircraft in No. 6, what is its speed?
9. For the aircraft in No. 6, what is its heading?
10. For the aircraft in No. 6, what is its altitude?
11. For the aircraft in No. 6, what is its type?
12. For the aircraft in No. 6, what is its frequency?
13. For the aircraft in No. 6, what is its callsign?
14. Is weather an impact on its route?
15. Are voice communications okay on its route?
16. How much time is available on fuel remaining?
17. Are there any hazards or emergencies that will affect performance?
18. Is the aircraft within 5 miles of its airspace boundaries?
19. Distance to inner ADIZ?
20. What is the range and bearing from the mission aircraft to the target aircraft?
21. Where is the intercept point?
22. How much time to intercept?
23. Is an intercept over water possible?
24. What is the mission type?

Note. ID = identification; ADIZ = Air Defense Identification Zone.

Complacency

Complacency is a hypothetical construct that is used to explain a failure to monitor a highly reliable automated system effectively. The suggestion is that because it is perceived as highly reliable, operators may not merely trust it, but may trust it too much. Moray and Inagaki (2000) have argued that there is little evidence for complacent behavior. They further argue that its existence cannot be established without prior specification of optimal behavior as a benchmark and that previous research has not done that. In their view, complacency is concerned with attention (monitoring, sampling) and not with detection as much of the research tends to imply. At first glance, this may seem like a reincarnation of the classical `vigilance' decrement, but there is an important distinction. Vigilance decrements are associated with very simple signals, such as those generated by radar or sonar systems. Although there has been some discussion of classical vigilance decrements in the supervisory control of complex systems, there is little evidence that they are of importance (Moray & Haudegond, 1998). Further, the conceptualization of the underlying cause is different; complacency emerges from excessive trust while vigilance decrements emerge from something like fatigue or habituation.

Dimensions of Supervisory Control: Summary

The guidance taken from the project objective was that the scale to be developed should be subjective and address issues of workload, situation awareness, and complacency. Further, it was desirable that the scale content not have to be reconstructed for different tasks.

The NASA TLX is such a scale which has the benefit of being reliable, valid, sensitive, and easy to administer, although it addresses only one of the three factors of concern in this project. Additionally, the project objective called for a rating instrument somewhat like that of the Cooper-Harper scale. Accordingly, we have developed scales of that form for assessment of workload, situation awareness, and complacency. Within our discussion of workload instruments, the six NASA TLX scale factors (mental, physical, temporal, effort, performance, and frustration) and the three SWAT scale factors (time load, mental effort, and psychological stress) were noted. Although the terminology between the two sets of factors is slightly different, we concluded that mental load and temporal load are common factors for the two scales. This led to a decision to develop separate rating scales for each of these.

We used a review of situation awareness to confirm our understanding of the nature of that construct, but the specific instruments commonly used require the content to be developed for each specific domain. While we used the content of Tables 2 and 3 to guide our understanding, we developed a Cooper-Harper type instrument for assessment of situational awareness that did not require that sort of content.

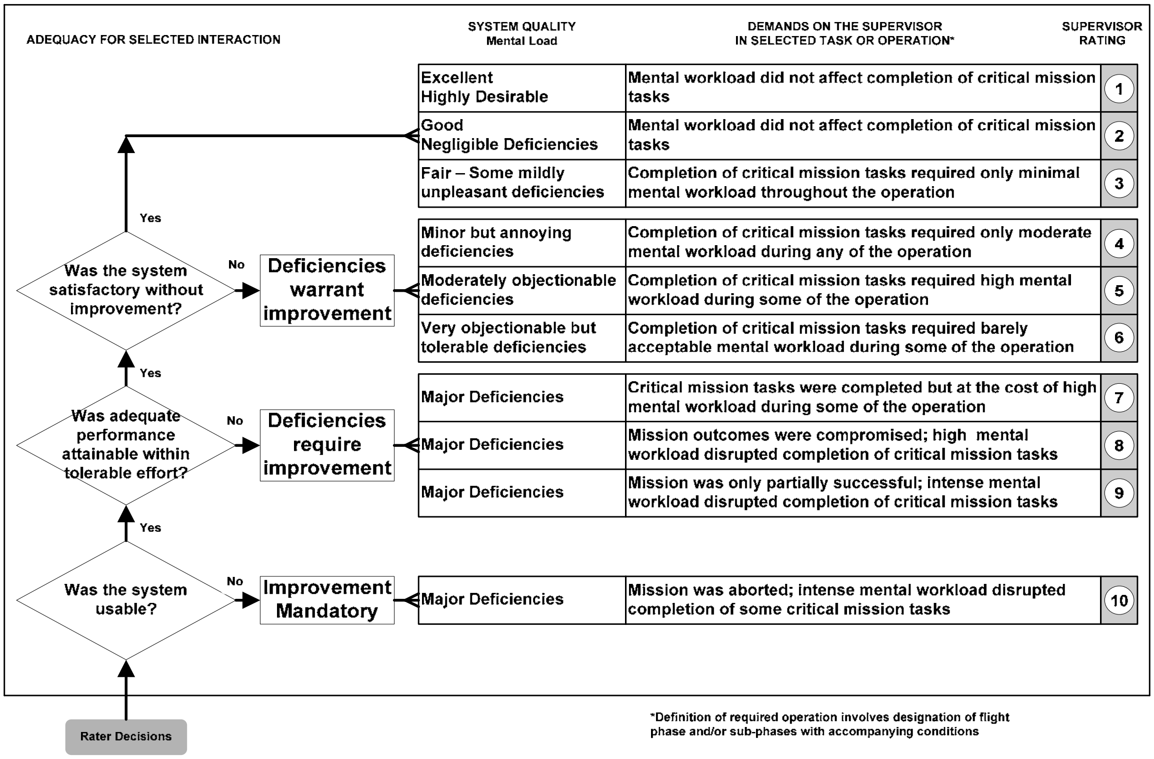
Moray and Inagaki (2000) suggested that the development of a scale for assessment of complacency might be a fruitless exercise. Nevertheless, in light of the project objective and reflecting on the possibility that this might not be the last word on this particular construct, we used their views as a guide in constructing a Cooper-Harper type instrument for assessment of complacency. Specifically, we prefer the term trust rather than complacency because it is more concrete and oriented the scale questions towards assessment of attitudes about effectiveness of attention processes rather than of detection processes.

Finally, we believe it a somewhat superficial attitude towards diagnosis to view a scale as diagnostic because it assesses multiple dimensions. We suggest that diagnosis requires identification of the specific system feature that is causing the problem. To that end, we developed a supplementary scale based on the structure of Rasmussen's decision ladder (Rasmussen, 1986) that can be used for diagnosis. The general rating strategy was to apply the Cooper-Harper style of scales in a first pass and then, if the ratings indicated problems with the system and the project goals warranted, the decision-ladder scales would be administered in one or more successive passes. Those decision ladder scales explore visibility and usability as discussed by Woods and his colleagues (e.g., Woods & Hollnagel, 2006) and also functionality as we have discussed above.

Supervisory Control Rating Scales

Cooper-Harper Style Scales

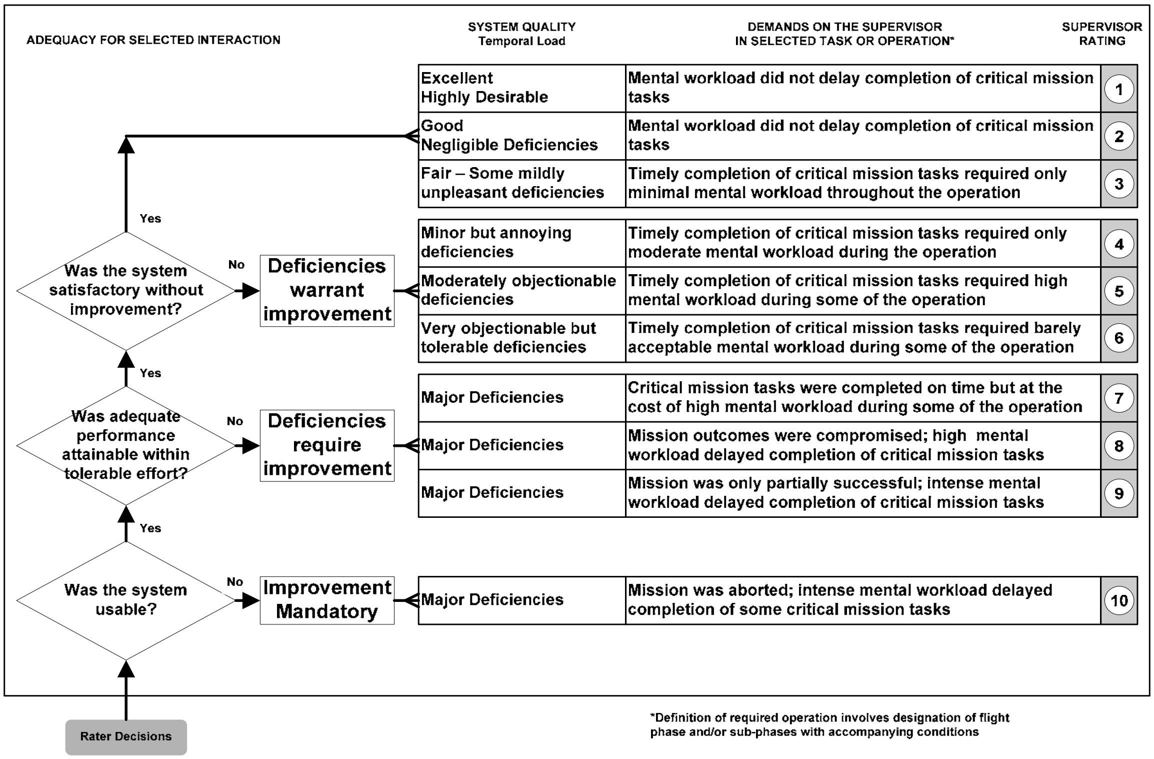
Scales for mental load, temporal load, situation awareness, and trust are shown in Figures 4 to 7. As for the Cooper-Harper instrument, the scale for each of these four instruments ranges from 1 to 10, with 1 indicating the best handling characteristics and 10 the worst. The rater begins at the bottom and continues upwards unless the decision point leads to a negative response. A negative response directs the rater to the right to identify a specific rating to be assigned.


ADEQUACY FOR SELECTED INTERACTION
Yes
Was the system satisfactory without improvement?
No
Deficiencies warrant improvement
Yes
Was adequate performance attainable within tolerable effort?
No
Deficiencies require improvement
Yes
Was the system usable?
No
Improvement Mandatory
Rater Decisions

SYSTEM QUALITY Mental Load	DEMANDS ON THE SUPERVISOR IN SELECTED TASK OR OPERATION*	SUPERVISOR RATING
Excellent Highly Desirable	Mental workload did not affect completion of critical mission tasks	1
Good Negligible Deficiencies	Mental workload did not affect completion of critical mission tasks	2
Fair – Some mildly unpleasant deficiencies	Completion of critical mission tasks required only minimal mental workload throughout the operation	3
Minor but annoying deficiencies	Completion of critical mission tasks required only moderate mental workload during any of the operation	4
Moderately objectionable deficiencies	Completion of critical mission tasks required high mental workload during some of the operation	5
Very objectionable but tolerable deficiencies	Completion of critical mission tasks required barely acceptable mental workload during some of the operation	6
Major Deficiencies	Critical mission tasks were completed but at the cost of high mental workload during some of the operation	7
Major Deficiencies	Mission outcomes were compromised; high mental workload disrupted completion of critical mission tasks	8
Major Deficiencies	Mission was only partially successful; intense mental workload disrupted completion of critical mission tasks	9
Major Deficiencies	Mission was aborted; intense mental workload disrupted completion of some critical mission tasks	10

*Definition of required operation involves designation of flight phase and/or sub-phases with accompanying conditions

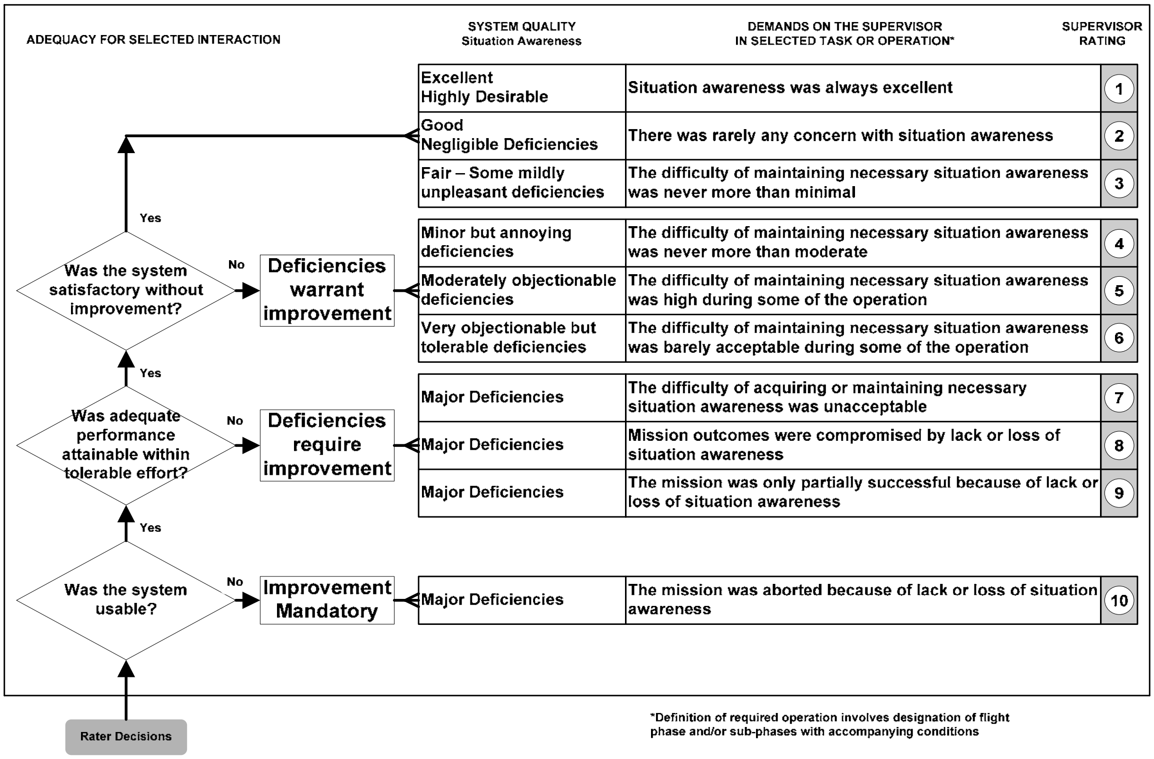
Figure 4. A Cooper-Harper style 10-point rating scale for mental load.


ADEQUACY FOR SELECTED INTERACTION
Rater Decisions
Was the system usable?
No
Improvement Mandatory
Yes
Was adequate performance attainable within tolerable effort?
No
Deficiencies require improvement
Yes
Was the system satisfactory without improvement?
No
Deficiencies warrant improvement
Yes

SYSTEM QUALITY Temporal Load	DEMANDS ON THE SUPERVISOR IN SELECTED TASK OR OPERATION*	SUPERVISOR RATING
Excellent Highly Desirable	Mental workload did not delay completion of critical mission tasks	1
Good Negligible Deficiencies	Mental workload did not delay completion of critical mission tasks	2
Fair – Some mildly unpleasant deficiencies	Timely completion of critical mission tasks required only minimal mental workload throughout the operation	3
Minor but annoying deficiencies	Timely completion of critical mission tasks required only moderate mental workload during the operation	4
Moderately objectionable deficiencies	Timely completion of critical mission tasks required high mental workload during some of the operation	5
Very objectionable but tolerable deficiencies	Timely completion of critical mission tasks required barely acceptable mental workload during some of the operation	6
Major Deficiencies	Critical mission tasks were completed on time but at the cost of high mental workload during some of the operation	7
Major Deficiencies	Mission outcomes were compromised; high mental workload delayed completion of critical mission tasks	8
Major Deficiencies	Mission was only partially successful; intense mental workload delayed completion of critical mission tasks	9
Major Deficiencies	Mission was aborted; intense mental workload delayed completion of some critical mission tasks	10

*Definition of required operation involves designation of flight phase and/or sub-phases with accompanying conditions

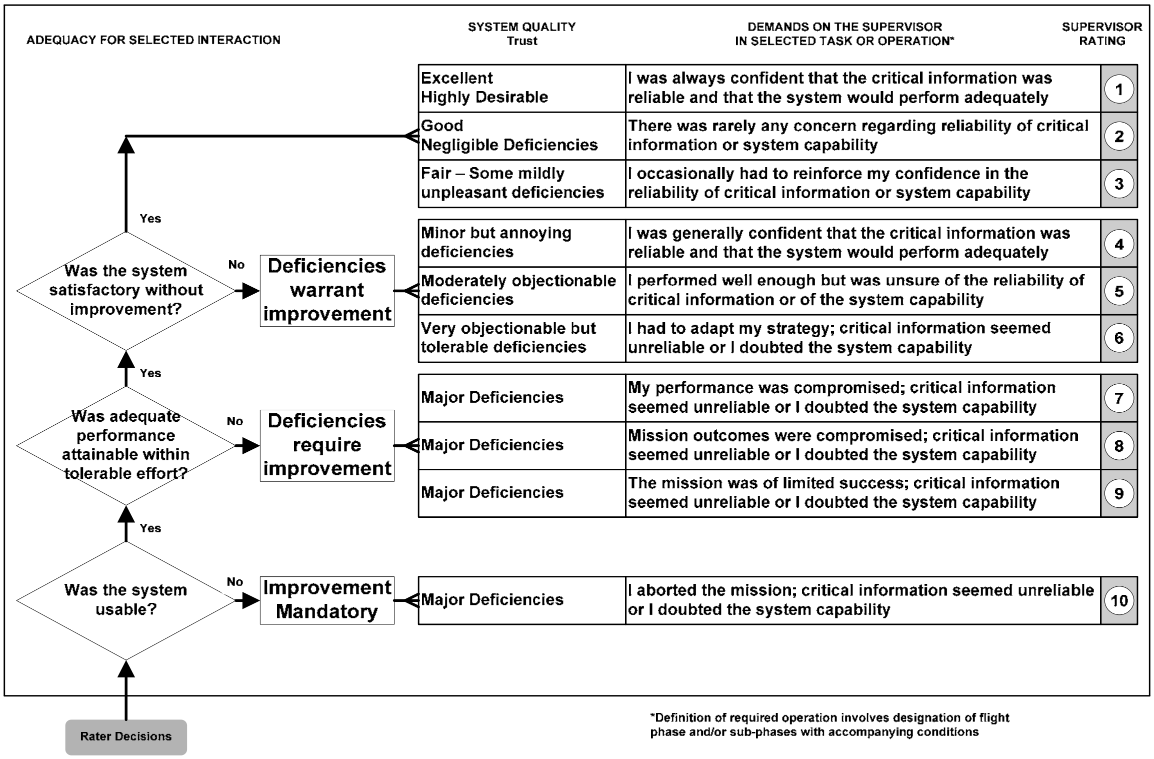
Figure 5. A Cooper-Harper style 10-point rating scale for temporal load.


ADEQUACY FOR SELECTED INTERACTION
Rater Decisions
Was the system usable?
No
Improvement Mandatory
Yes
Was adequate performance attainable within tolerable effort?
No
Deficiencies require improvement
Yes
Was the system satisfactory without improvement?
No
Deficiencies warrant improvement
Yes

SYSTEM QUALITY Situation Awareness	DEMANDS ON THE SUPERVISOR IN SELECTED TASK OR OPERATION*	SUPERVISOR RATING
Excellent Highly Desirable	Situation awareness was always excellent	1
Good Negligible Deficiencies	There was rarely any concern with situation awareness	2
Fair – Some mildly unpleasant deficiencies	The difficulty of maintaining necessary situation awareness was never more than minimal	3
Minor but annoying deficiencies	The difficulty of maintaining necessary situation awareness was never more than moderate	4
Moderately objectionable deficiencies	The difficulty of maintaining necessary situation awareness was high during some of the operation	5
Very objectionable but tolerable deficiencies	The difficulty of maintaining necessary situation awareness was barely acceptable during some of the operation	6
Major Deficiencies	The difficulty of acquiring or maintaining necessary situation awareness was unacceptable	7
Major Deficiencies	Mission outcomes were compromised by lack or loss of situation awareness	8
Major Deficiencies	The mission was only partially successful because of lack or loss of situation awareness	9
Major Deficiencies	The mission was aborted because of lack or loss of situation awareness	10

*Definition of required operation involves designation of flight phase and/or sub-phases with accompanying conditions

Figure 6. A Cooper-Harper style 10-point rating scale for situation awareness.


ADEQUACY FOR SELECTED INTERACTION
SYSTEM QUALITY
Trust
DEMANDS ON THE SUPERVISOR
IN SELECTED TASK OR OPERATION*
SUPERVISOR
RATING
Excellent
Highly Desirable
I was always confident that the critical information was reliable and that the system would perform adequately
1
Good
Negligible Deficiencies
There was rarely any concern regarding reliability of critical information or system capability
2
Fair – Some mildly unpleasant deficiencies
I occasionally had to reinforce my confidence in the reliability of critical information or system capability
3
Minor but annoying deficiencies
I was generally confident that the critical information was reliable and that the system would perform adequately
4
Moderately objectionable deficiencies
I performed well enough but was unsure of the reliability of critical information or of the system capability
5
Very objectionable but tolerable deficiencies
I had to adapt my strategy; critical information seemed unreliable or I doubted the system capability
6
Major Deficiencies
My performance was compromised; critical information seemed unreliable or I doubted the system capability
7
Major Deficiencies
Mission outcomes were compromised; critical information seemed unreliable or I doubted the system capability
8
Major Deficiencies
The mission was of limited success; critical information seemed unreliable or I doubted the system capability
9
Major Deficiencies
I aborted the mission; critical information seemed unreliable or I doubted the system capability
10
Yes
Was the system satisfactory without improvement?
No
Deficiencies warrant improvement
Yes
Was adequate performance attainable within tolerable effort?
No
Deficiencies require improvement
Yes
Was the system usable?
No
Improvement Mandatory
Rater Decisions

*Definition of required operation involves designation of flight phase and/or sub-phases with accompanying conditions

Figure 7. A Cooper-Harper style 10-point rating scale for trust.

As noted previously, a Cooper-Harper style rating scale is not diagnostic. It would be possible to follow the strategy that is commonly used in application of the Cooper-Harper scale to flight control assessment, that being to have the rater provide comments in conjunction with the ratings that would be more diagnostic. Without negating the potential of that approach to assessment of supervisory control, here we outline a different and more systematic approach based on Rasmussen's Decision Ladder (Rasmussen, 1986).

Decision-Ladder Scales

The Decision-Ladder scale was developed with supervisory control of an unmanned aerial vehicle (UAV) mission in mind. It will presumably have to be adjusted if it is to be applied to a different type of supervisory control task.

The left panel of Figure 8 depicts our form of the Decision Ladder. It was developed by editing and reformatting the Decision Ladder as first depicted by Rasmussen (1986) to clarify some of the confusions about the original form. However, our editing and reformatting have not, in any way, changed the underlying constructs.

In normal use, the Decision Ladder is used to map out the cognitive states and processes for a work problem or one of its components. Note that it does not represent a set trajectory for a control task; workers are flexible and will generate different trajectories at different times. It maps cognitive states and processes that might be used in supervisory control and therefore shows what cognitive states and processes should be targeted in a redesign effort. In our development of it for use as a rating scale, it indicates what cognitive states and processes are at the basis of any challenges to effective supervisory control. In that respect, it offers a more fine-grained diagnostic tool than the NASA TLX or the SWAT.

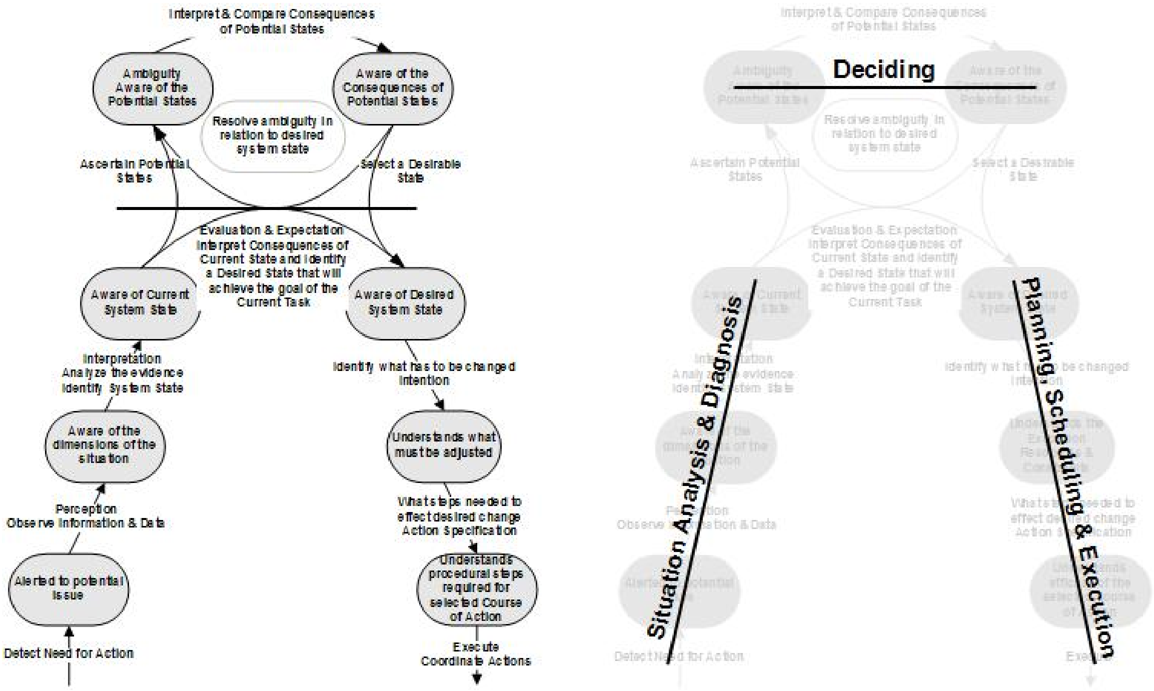
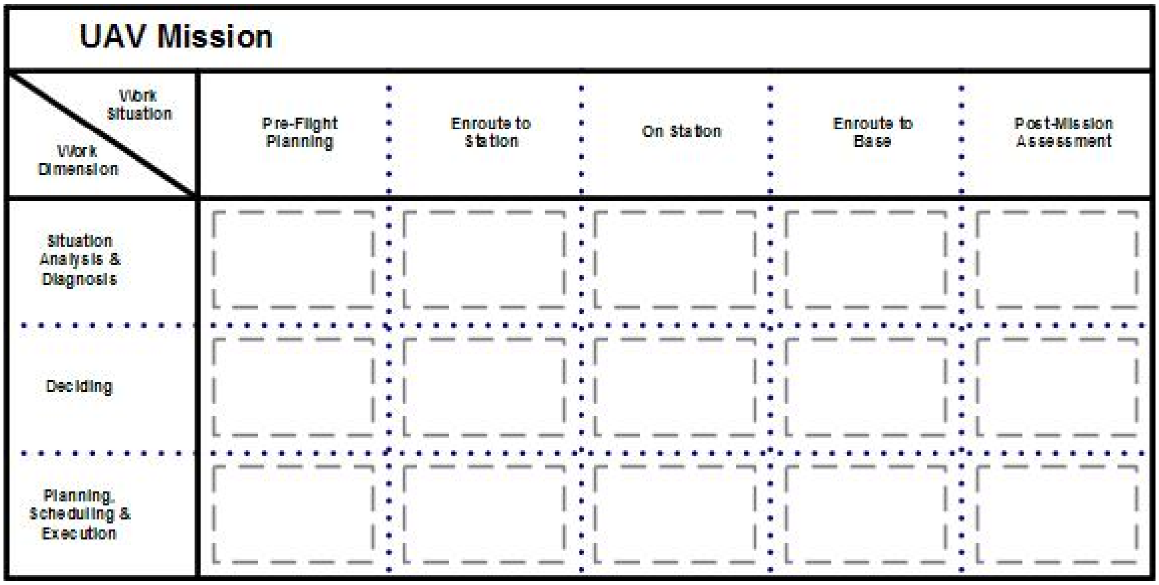

Interpret & Compare Consequences of Potential States
Ambiguity Aware of the Potential States
Aware of the Consequences of Potential States
Resolve ambiguity in relation to desired system state
Ascertain Potential States
Select a Desirable State
Evaluation & Expectation Interpret Consequences of Current State and Identify a Desired State that will achieve the goal of the Current Task
Aware of Current System State
Aware of Desired System State
Interpretation Analyze the evidence Identify System State
Identify what has to be changed Intention
Aware of the dimensions of the situation
Understands what must be adjusted
Perception Observe Information & Data
What steps needed to effect desired change Action Specification
Alerted to potential issue
Understands procedural steps required for selected Course of Action
Detect Need for Action
Execute Coordinate Actions
Deciding
Situation Analysis & Diagnosis
Planning, Scheduling & Execution

Figure 8. A Decision Ladder (left panel) inspired by but modified from Rasmussen (1986) and the Decision Ladder's three main stages (right panel).

As shown in Figure 8 (right panel) a Decision Ladder identifies three dimensions of supervisory control:

- Situation Analysis and Diagnosis: detect a need for action, observe all essential information and data, analyze the evidence, and interpret it to understand the situation
- Decision: anticipate the consequences of the current situation and understand the potential outcomes of different decisions
- Planning, Scheduling, and Execution: plan a course of action to achieve the task goal and understand how to manipulate the system to execute that course of action

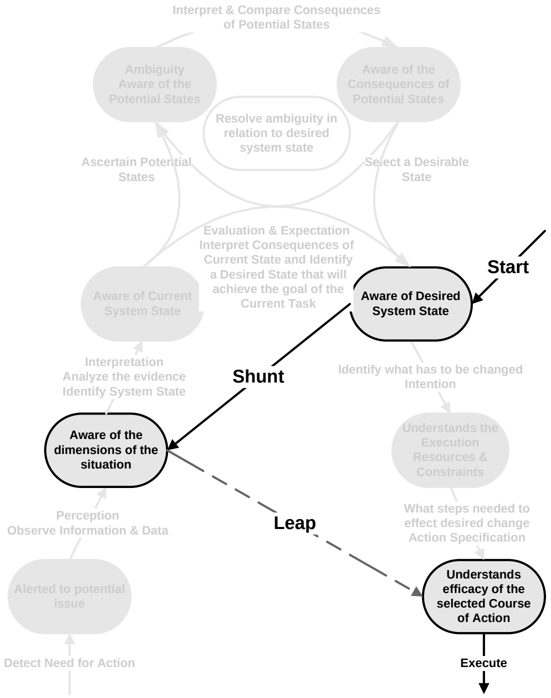
In a first pass with the Decision-Ladder instrument, raters would be asked to assess the capability of the system for each phase of the mission on each of these dimensions. Figure 9 depicts a contextual activity map for collection of those ratings. The mission phases are shown horizontally across the top and the three decision ladder dimensions vertically along the left.



UAV Mission

Work Situation / Work Dimension	Pre-Flight Planning	Enroute to Station	On Station	Enroute to Base	Post-Mission Assessment
Situation Analysis & Diagnosis					
Deciding					
Planning, Scheduling & Execution					

Figure 9. A contextual activity rating scale for an unmanned air vehicle mission.



Interpret & Compare Consequences of Potential States
Ambiguity Aware of the Potential States
Aware of the Consequences of Potential States
Resolve ambiguity in relation to desired system state
Ascertain Potential States
Select a Desirable State
Evaluation & Expectation Interpret Consequences of Current State and Identify a Desired State that will achieve the goal of the Current Task
Aware of Current System State
Aware of Desired System State
Start
Shunt
Interpretation Analyze the evidence Identify System State
Identify what has to be changed Intention
Aware of the dimensions of the situation
Understands the Execution Resources & Constraints
Perception Observe Information & Data
Leap
What steps needed to effect desired change Action Specification
Alerted to potential issue
Understands efficacy of the selected Course of Action
Detect Need for Action
Execute

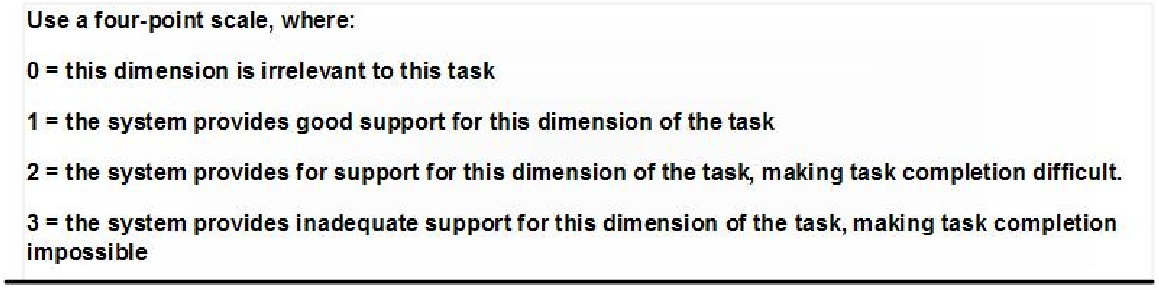
Figure 10. A simple control task as mapped onto the decision ladder.

If the diagnostic information made available from this first pass through the Decision Ladder is thought to be insufficiently fine-grained to be fully diagnostic, a second pass that focuses on individual states and processes as represented in the full Decision Ladder (Figure 8, left panel) could be undertaken. This would require a narrative that specification of the supervisory control task that is to be assessed, which would then have to be mapped onto a decision ladder as illustrated in Figure 10.

The task narrative represented in Figure 10 is based on a description of how an expert targeteer scheduled munitions for an attack on a Time Sensitive Target (TST) during his tour in Iraq. As shown in Figure 10, this targeteer was initially aware of the desired system state. As described by the targeteer, he almost always scheduled a particular type of munitions for a TST and typically did not explore other possibilities. This decision was based on his belief that these munitions were reliable and accurate and could be delivered with minimal pilot workload. As narrated by the targeteer, he would confirm that these munitions were suitable for this particular target situation, which is represented in Figure 10 by the shunt (an explicit cognitive process) from ”Aware of Desired System State” to “Aware of the Dimensions of the Situation”. Typically, as he reported, these munitions were suitable for the particular target situation and so he would conclude that he understood the efficacy of his selected course of action and would then insert his decision into the attack plan. These final steps are represented in Figure 10 by the leap (an implicit cognitive process) from “Aware of the Dimensions of the Situation” to “Understands Efficacy of the Selected Course of Action” and “Execute”.

This second pass through the decision ladder would focus only on the cognitive states and cognitive processes involved in the narrative. We conceived of two possibilities for rating this second pass, one involving a single dimension of judgment and the other involving two dimensions of judgment. The first of these is described in Table 4 and the second in Table 5.

Table 4. A Single-Dimension Assessment Strategy for a Specific Decision Ladder Trajectory



Use a four-point scale, where:

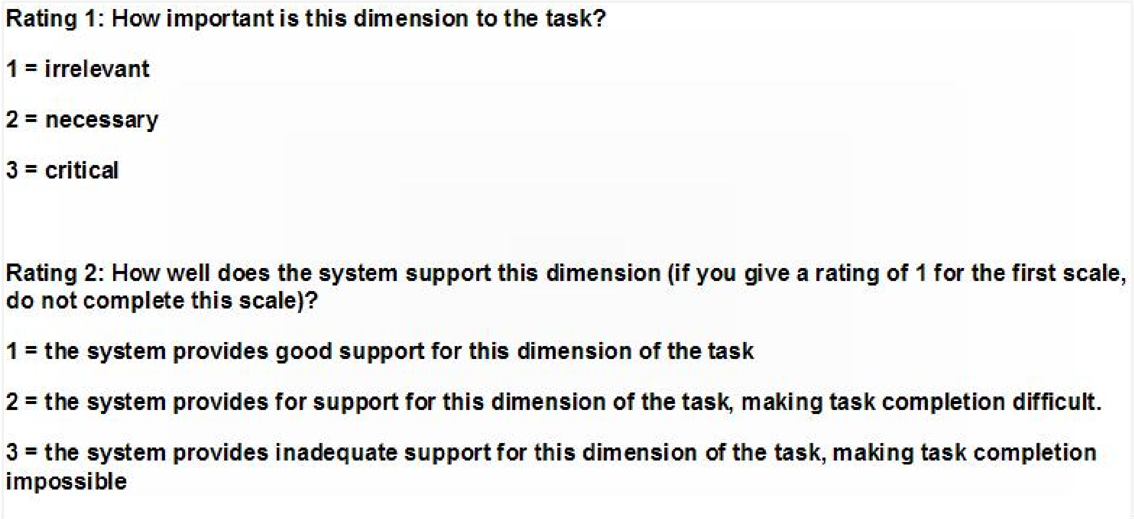
0 = this dimension is irrelevant to this task

1 = the system provides good support for this dimension of the task

2 = the system provides for support for this dimension of the task, making task completion difficult.

3 = the system provides inadequate support for this dimension of the task, making task completion impossible

Table 5. A Dual-Dimension Assessment Strategy for a Specific Decision Ladder Trajectory



Rating 1: How important is this dimension to the task?

1 = irrelevant

2 = necessary

3 = critical

Rating 2: How well does the system support this dimension (if you give a rating of 1 for the first scale, do not complete this scale)?

1 = the system provides good support for this dimension of the task

2 = the system provides for support for this dimension of the task, making task completion difficult.

3 = the system provides inadequate support for this dimension of the task, making task completion impossible

Conclusion: Evaluation of Rating Scales

We understand that the Air Force Research Laboratory intends to deploy this scale within an UAV evaluation scenario. While that evaluation could be planned at different levels of detail, a complete and systematic evaluation that would establish one or more components of the scales we have developed here would require evaluation of reliability, sensitivity, and validity. This is a substantial undertaking that would require collection of ratings from a number of subjects and detailed statistical analyses that would involve comparisons between scale measures and objective task performance measures to evaluate their psychometric characteristics, including sensitivity, internal consistency, test-retest reliability, and content validity.

References

Birmingham, H, P., & Taylor, F, V. (1954). A design philosophy for man-machine control systems. *Proceedings of the Institute of Radio Engineers, 42,* 1748-1758

Christoffersen, K. & Woods, D. D. (2002). How to make automated systems team players. *Advances in Human Performance and Cognitive Engineering Research, 2,* 1-12. Elsevier Science Ltd.

Cummings, M. L., Meyers, K. & Scott, S. D. (2006). Modified Cooper-Harper Evaluation Tool for Unmanned Vehicle Displays. *Fourth Annual Conference of UVS Canada,* Montebello, Quebec.

Dekker, S. W. A., & Woods, D. D. (2002). MABA-MABA or Abracadabra*? Progress on Human -Automation Co-ordination Cognition, Technology & Work, 4,* 240 -244.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 97–101). Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37,* 32–64.

European Organisation for the Safety of Air Navigation (2003*). Review of Workload Measurement, Analysis and Interpretation Methods.* Report # CARE-Integra-TRS-130-02-WP2.

Field, E. J. (1995). *Flying qualities of transport aircraft: Precognitive or compensatory?* Unpublished doctoral dissertation, College of Aeronautics, Cranfield University, Cranfield, England.

Harper, R. P., & Cooper, G. E. (1986). Handling qualities and pilot evaluation. *Journal of Guidance, Control and Dynamics, 9,* 515–530.

Harris, D., Gautrey, J., Payne, K., & Bailey, R. (2000). The Cranfield Aircraft Handling Qualities Rating Scale: A multidimensional approach to the assessment of aircraft handling qualities. *Aeronautical Journal, 104*, 191–198.

Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.)*, Human mental workload* (pp. 139–183). Amsterdam: North-Holland.

Hollnagel, E., & Woods, D. D. (2005). Joint Cognitive Systems: Patterns in Cognitive Systems Engineering. Boca Rotan, FL: CRC Press. ISBN: 0849328217.

Inagaki, T. (2003). Automation and the cost of authority. *International Journal of Industrial Ergonomics 31,* 169 –174.

Jones, D. G., & Endsley, M. R. (2004). Use of Real-Time Probes for Measuring Situation Awareness. *International Journal of Aviation Psychology, 14 (4),* 343 –367

Lintern, G. (2007). What is a Cognitive System? *Proceedings of the Fourteenth International Symposium on Aviation Psychology, (pp. 398-402).* Dayton, OH.

Miller, G. A. (1956). Information and memory. *Scientific American, 195,* 42-46.

Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomic Science, 1, (4),* 354-365.

Moray, N., & Haudegond, S. (1998). An absence of vigilance decrement in a complex dynamic task. *Proceedings of Annual Meeting of Human Factors and Ergonomics Society*, Chicago, IL.

Parasuraman, R., Sheridan T., & Wickens, C. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 30,* pp. 286-297.

Payne, K., & Harris, D. (2000). A Psychometric Approach to the Development of a Multidimensional Scale to Assess Aircraft Handling Qualities. *International Journal of Aviation Psychology, 10 (4),* 343-362.

Rasmussen, J. (1986). *Information processing and human machine interaction: an approach to cognitive engineering.* NY: Science Publishing, North Holland series in system science and engineering, 12.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N.

Meshkati (Eds.), *Human mental workload* (pp. 185–218). Amsterdam: North-Holland.

Sarter, N. B., & Woods, D. D. (1992). Pilot interaction with cockpit automation: Operational experiences with the flight management system. *International Journal of Aviation Psychology, 2 (4),* 303-321.

Sarter, N. B., & Woods, D. D. (1994). Pilot interaction with cockpit automation II: Operational experiences with the flight management system. *International Journal of Aviation Psychology, 4,* 1-28.

Sarter, N. B., Woods, D. D., & Billings, C. (1997). Automation Surprises. In G. Salvendy (Ed.), *Handbook of Human Factors/Ergonomics* (2nd ed.) (pp. 1926-1943), NY: Wiley.

Sheridan, T. B. (1988). Task Allocation and Supervisory Control. In M. Helander (ed.), *Handbook of Human-Computer Interaction* (pp. 159-173). Amsterdam: Elsevier Science Publishers.

Sheridan, T. B. (1997). Task Analysis, Task Allocation and Supervisory Control. In M. Helander, T.K. Landauer, P. Prabhu (eds.), *Handbook of Human-Computer Interaction, second completely revised edition* (pp. 87-105). Amsterdam: Elsevier Science Publishers.

Weick, K. E., & Sutcliffe, K. M. (2001). *Managing the unexpected: assuring high performance in an age of complexity.* San Francisco: John Wiley.

Wilson, D. J, & Riley, D. R. (1989). *Cooper–Harper rating variability.* Paper presented at the American Institute of Aeronautics and Astronautics Atmospheric Flight Mechanics Conference, Boston, MA.

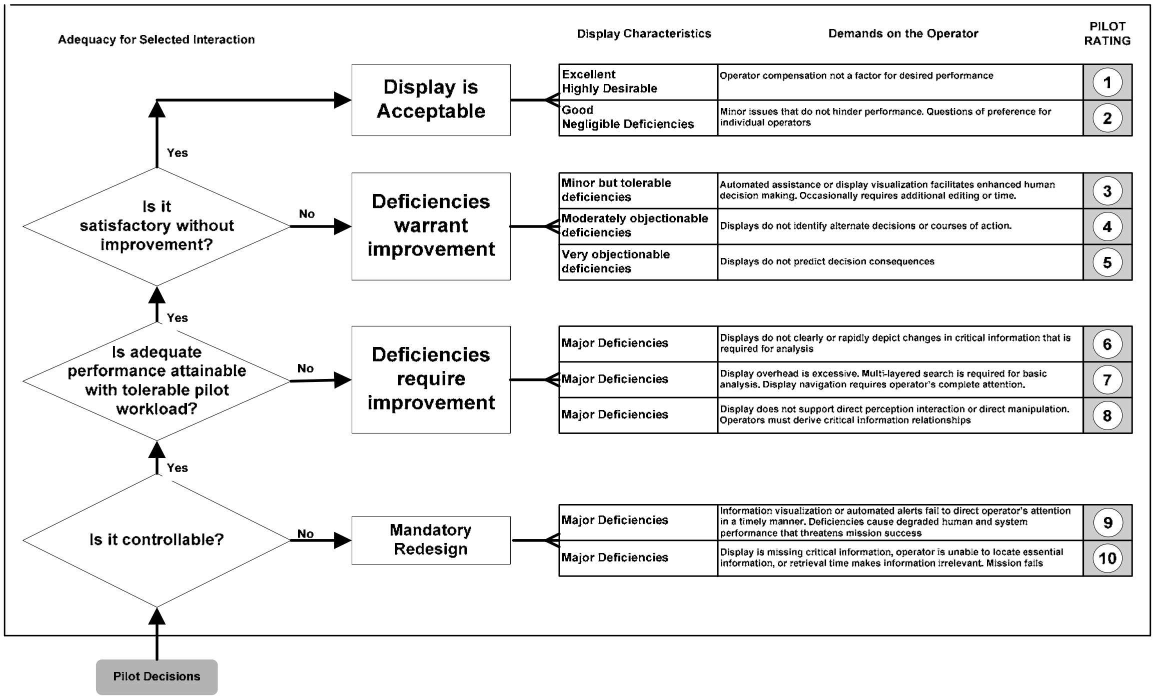
Woods, D. D., & Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering.* Boca Rotan, FL: CRC Press.

Woods, D. D., & Patterson, E. S. (2000). How Unexpected Events Produce an Escalation of Cognitive and Coordinative Demands. In P. A. Hancock & P. Desmond (Eds.), *Stress Workload and Fatigue.* Hillsdale, NJ: Lawrence Erlbaum.

Appendix A

Summary of Cummings, Meyers, and Scott (2006)

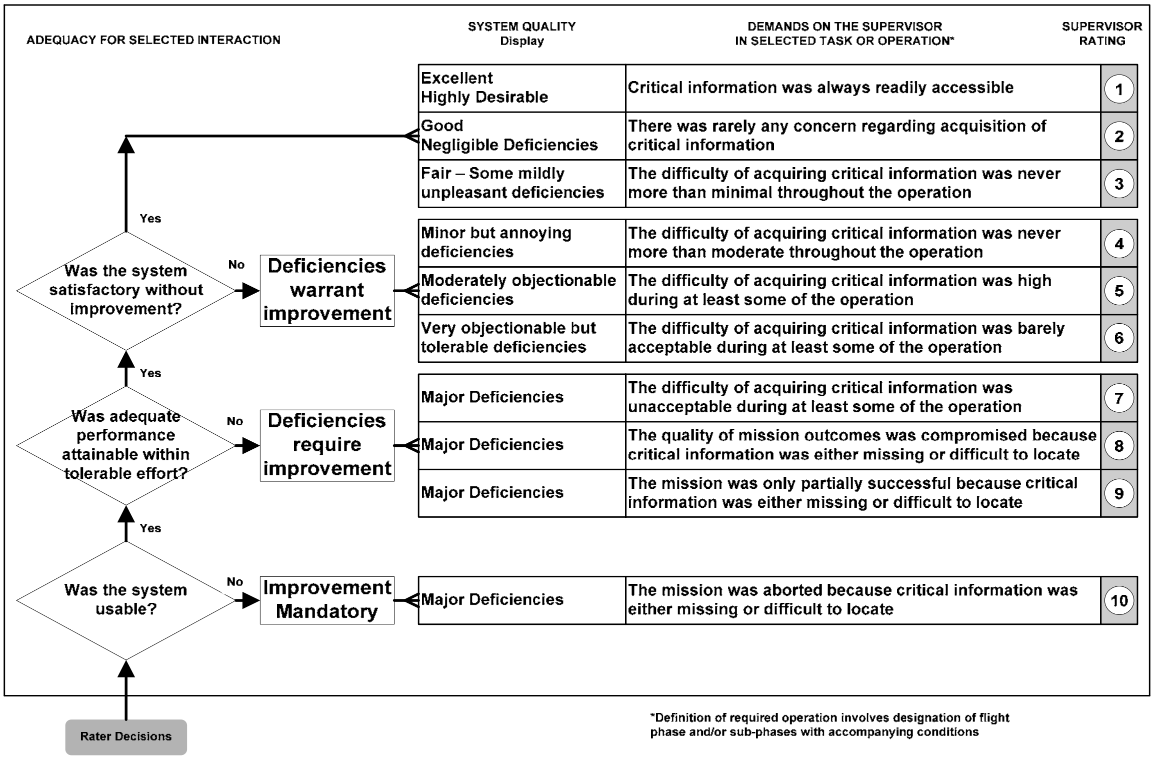
We have reviewed an adaptation of the Cooper-Harper scale described by Cummings et al. (M. L. Cummings, K. Meyers, and S. D. Scott, "*Modified Cooper-Harper Evaluation Tool for Unmanned Vehicle Displays*," presented at the UVS Canada Conference, 2006). This in modification of the Cooper-Harper scale is specific to the usefulness of displays and was evaluated against a display, developed within the Humans and Automation Laboratory, Massachusetts Institute of Technology, for UAV control. We have duplicated that scale here as Figure 11. One concern that we have with this scale is that the ratings given to the display on the scale by the different experimental subjects vary widely, with some subjects evaluating the display as very good and others evaluating it as very poor. We remain uncertain regarding the cause of such diversity, but speculate that it occurred because the experimental subjects were not given an opportunity to benchmark their ratings.


Adequacy for Selected Interaction
Pilot Decisions
Is it controllable?
No
Mandatory Redesign
Yes
Is adequate performance attainable with tolerable pilot workload?
No
Deficiencies require improvement
Yes
Is it satisfactory without improvement?
No
Deficiencies warrant improvement
Yes
Display is Acceptable

Display Characteristics	Demands on the Operator	PILOT RATING
Excellent Highly Desirable	Operator compensation not a factor for desired performance	1
Good Negligible Deficiencies	Minor issues that do not hinder performance. Questions of preference for individual operators	2
Minor but tolerable deficiencies	Automated assistance or display visualization facilitates enhanced human decision making. Occasionally requires additional editing or time.	3
Moderately objectionable deficiencies	Displays do not identify alternate decisions or courses of action.	4
Very objectionable deficiencies	Displays do not predict decision consequences	5
Major Deficiencies	Displays do not clearly or rapidly depict changes in critical information that is required for analysis	6
Major Deficiencies	Display overhead is excessive. Multi-layered search is required for basic analysis. Display navigation requires operator's complete attention.	7
Major Deficiencies	Display does not support direct perception interaction or direct manipulation. Operators must derive critical information relationships	8
Major Deficiencies	Information visualization or automated alerts fail to direct operator's attention in a timely manner. Deficiencies cause degraded human and system performance that threatens mission success	9
Major Deficiencies	Display is missing critical information, operator is unable to locate essential information, or retrieval time makes information irrelevant. Mission fails	10

Figure 11. A Cooper-Harper style of rating scale for display evaluation as developed by Cummings, Meyers, and Scott (2006).

A further concern we have with the Cummings et al (2006) scale is that the different levels appear to be tapping different qualities. The original Cooper-Harper scale tapped only one quality, that being aircraft controllability. In contrast, Cummings et al. have developed a scale in which ratings of 1 and 2 assess an unspecified quality, ratings of 3 to 5 assess decision making and assessment, ratings of 6 to 8 assess the form and structure of information, and ratings of 9 and 10 assess availability of information. We remain concerned that such deviations from the accepted structure of a Cooper-Harper style scale invalidate the connection to the background work that stimulated our interest in it. As a contrasting alternative to the display evaluation scale developed by Cummings et al (2006), we have developed a display evaluation that consistently evaluates accessibility to and availability of information across all levels of rating. That scale is shown in Figure A-2.


ADEQUACY FOR SELECTED INTERACTION
Was the system satisfactory without improvement?
Yes
No
Deficiencies warrant improvement
Was adequate performance attainable within tolerable effort?
Yes
No
Deficiencies require improvement
Was the system usable?
Yes
No
Improvement Mandatory
Rater Decisions

SYSTEM QUALITY Display	DEMANDS ON THE SUPERVISOR IN SELECTED TASK OR OPERATION*	SUPERVISOR RATING
Excellent Highly Desirable	Critical information was always readily accessible	1
Good Negligible Deficiencies	There was rarely any concern regarding acquisition of critical information	2
Fair – Some mildly unpleasant deficiencies	The difficulty of acquiring critical information was never more than minimal throughout the operation	3
Minor but annoying deficiencies	The difficulty of acquiring critical information was never more than moderate throughout the operation	4
Moderately objectionable deficiencies	The difficulty of acquiring critical information was high during at least some of the operation	5
Very objectionable but tolerable deficiencies	The difficulty of acquiring critical information was barely acceptable during at least some of the operation	6
Major Deficiencies	The difficulty of acquiring critical information was unacceptable during at least some of the operation	7
Major Deficiencies	The quality of mission outcomes was compromised because critical information was either missing or difficult to locate	8
Major Deficiencies	The mission was only partially successful because critical information was either missing or difficult to locate	9
Major Deficiencies	The mission was aborted because critical information was either missing or difficult to locate	10

*Definition of required operation involves designation of flight phase and/or sub-phases with accompanying conditions

Figure A-2. A Cooper-Harper style of rating scale for display evaluation that consistently evaluates accessibility to and availability of information across all levels of rating.