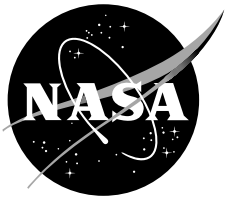
**NASA/TM–1999-208795**
**USAAMCOM-AFDD/TR-00-A-002**





# Situation Awareness and Workload Measures for SAFOR

*Joe De Maio and Sandra G. Hart*

**October 1999**

# The NASA STI Program Office . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the Lead Center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

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

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

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

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

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

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

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results . . . even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at *http://www.sti.nasa.gov*

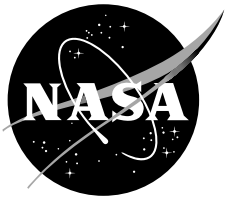
- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA Access Help Desk at (301) 621-0134

- Telephone the NASA Access Help Desk at (301) 621-0390

- Write to:
  NASA Access Help Desk
  NASA Center for AeroSpace Information
  7121 Standard Drive
  Hanover, MD 21076-1320

**NASA/TM–1999-208795**
**USAAMCOM-AFDD/TR-00-A-002**





# Situation Awareness and Workload Measures for SAFOR

*Joe De Maio*
*Aeroflightdynamics Directorate*
*U.S. Army Aviation and Missile Command*
*Ames Research Center, Moffett Field, California*

*Sandra G. Hart*
*Ames Research Center, Moffett Field, California*

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035-1000

**October 1999**

Available from:

NASA Center for AeroSpace Information
7121 Standard Drive
Hanover, MD 21076-1320
(301) 621-0390

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
(703) 487-4650

SUMMARY

The present research was performed as part of the U. S. Army Human Systems Integration and the NASA Safe All-weather Flight Operations for Rotorcraft (SAFOR) programs. The purpose of the work was to investigate the utility of two measurement tools developed by the British Defense Evaluation Research Agency. These tools were a subjective workload assessment scale, the DRA Workload Scale (DRAWS) and a situation awareness measurement tool in which the crews self-evaluation of performance is compared against actual performance in order to determine what information the crew attended to during the performance. These two measurement tools were evaluated in the context of a test of innovation approach to alerting the crew by way of a helmet mounted display. The DRAWS was found to be usable, but it offered no advantages over extant scales, and it had only limited resolution. The performance self-evaluation metric of situation awareness was found to be highly effective.

# INTRODUCTION

The present research was performed as part of the U. S. Army Human Systems Integration and the NASA Safe All-weather Flight Operations for Rotorcraft (SAFOR) programs. The goal of our part of this program is to make dramatic reductions in the rate and severity of civil rotorcraft accidents. A major factor in pilot error is the failure to perceive flight related information or to interpret isolated facts to guide correct behavior. The term, situation awareness, includes this perception and interpretation of relevant information. Another factor that affects crew performance is workload. Workload and situation awareness are independent but interacting factors. The present research evaluated techniques for measuring both factors.

The present research was also shaped by a cooperative effort between the Army Aeroflightdynamics Directorate, located at NASA Ames Research Center and the British Defense Evaluation and Research Agency (DERA). The DERA is developing a battery of workload and situation awareness measures. These might be applicable to both U. S. and British research. The present research evaluated variants of two measurement techniques, the DRA Workload Scale (DRAWS) and self-assessment probes. DRAWS is a four-dimensional, subjective rating of perceived workload. The self-assessment probes are performance self-evaluations in which the individual rates his own performance against pre-determined criteria. The accuracy of the performance self-evaluation provides a measure of situation awareness.

There has been much discussion regarding the exact meaning of the term situation awareness (McMillan, Bushman, and Judge, 1996). Several approaches have been proposed to measure the construct. Fracker (1991) distinguished between two broad classes of situation awareness measures. Explicit measures are those that require the operator to recall facts relevant to the performance of the task, that is, the operator tells the evaluator explicitly what he knows about the task. This is contrasted with implicit measures. In implicit measurement, the evaluator measures task performance and infers a level of operator situation awareness from performance. The logic here is that if the operator were aware of a certain fact, he would perform a certain action. If he fails to perform the action, then he must have been unaware of the relevant fact.

There are difficulties associated with both implicit and explicit measurement of situation awareness. While the logic of implicit measurement is sound, that is, if A implies B, then not B implies not A, it is not easy to be certain that knowledge of a given fact will lead inevitably to a specific behavior, that is, that the premise is in fact true. The more complex the knowledge and the behavior, the more difficult it becomes to be certain of the linkage between them. As a result, implicit measures have been used for only simple facts and responses (e.g., Eubanks and Killeen, 1983). Even when the logical requirements of implicit measures are met, there can still be problems with underlying statistical assumptions. For example, Eubanks and Killeen assumed normality and equal variance of signal and non-signal in order to apply the theory of Signal Detectability to a targeting task. Long and Waag (1981) have pointed out that

this assumption is wrong for many supra-threshold tasks where situation awareness might be a concern.

While explicit measures do not require this sine qua non relationship with performance, the issue of relevance of facts remains. In explicit measurement, the evaluator queries the operator about specific facts determined a priori to be relevant to the task. The risk here is that the importance of a particular fact may vary over the course of task execution. Evaluations of situation awareness may differ depending on the part of the task queried using that fact (Fracker, op cit). For example, Endsley (1995b) used altitude as a query subject in an evaluation of the situation awareness global assessment technique (SAGAT). The flight task was a combat air patrol. This mission can be broken into two major components, orbit and air combat. During orbit the pilot is to maintain a constant altitude, and he monitors the altimeter continually. During air combat, maintaining a specific altitude is relatively unimportant, and the pilot focuses more on control variables to achieve a tactical advantage. As a result Endsley found very high variability, with evidence of good situation awareness (accurate recall of altitude) mixed with virtually non-existent situation awareness (errors over 20,000 ft). This risk may be minimized by using a battery of well chosen queries, from which inappropriate variables can be culled.

Explicit measures of situation awareness have proven most popular because of their versatility. Explicit measures can address not only raw facts (e.g., altitude), but also state concepts consisting of an aggregation of facts (e.g., tactical advantage) and also future states (e.g., engagement outcome). Endsley (1995a) has labeled raw facts, aggregated state concepts and future states situation awareness levels one, two, and three respectively.

Because explicit measures require the operator to respond to a query, there is always a memory component to the response. The memory component can be minimized by halting the task and querying the operator about his state immediately prior to the halt. This has been called a concurrent memory probe Fracker, op cit). The liability of the concurrent probe is that the task must be suspended or terminated to allow the query. This is not always feasible, and even when it is, the halt can severely disrupt the performance of the task.

Fracker (op cit) has suggested that level two and three factors may persist long enough to be probed retroactively, that is following normal task completion. Endsley (1995b) has tested the effect of delaying report by querying a number of facts following task halt. She has shown that even level one concurrent memory probes can be stable when many facts must be recalled.

The present research evaluated a form of retroactive, level two query. The self-evaluation required the pilot to integrate a number of facts about the task situation and his performance in order to produce an evaluation. Our queries differed from those generally applied. Whereas level two queries generally require the aggregation of a defined set of facts available to the operator, our self-evaluations queried the overall performance of a task without asking for values of specific variables at specific times. In some instances the ultimate performance data for grading the task was not even available to the pilot. We determined

what information the pilot attends to during task performance by comparing the self-evaluations with objective evaluations of performance.

A secondary part of the research was to examine the DRA Workload Scale (DRAWS), developed by the Defense Evaluation and Research Agency in the United Kingdom. The DERA plans to use DRAWS to evaluate workload in the Covert Night/Day Operations for Rotorcraft (CONDOR) program to develop a wide field of view, color, helmet-mounted display. DRAWS uses four subjective scales drawn from an information processing context. These are: Input, the workload associated with perceiving things; Central (Processing), the workload associated with interpreting information and deciding on an action; Output, the workload associated with overt action; and Time, the pressure to act quickly. DRAWS is very much focused on the performance of the task. Three of its four axes all relate to the process of performing an act itself rather than to feelings of the operator or characteristics of the work environment that are independent of the task.

The present evaluation of workload and situation awareness metrics was performed along with a test of helmet mounted display symbology. We tested two types of helmet display symbology, navigation aids and alerts. We present the results of this test elsewhere (De Maio and Hart, in preparation), but we present an overview of the display work below to provide a context for the metrics evaluation. The simulated mission tasking was designed around research issues associated with the symbology evaluation, and the workload and situation awareness data were gathered at appropriate times.

There were three navigation symbology conditions: an AH-64 pilot night vision system baseline, a course deviation indicator analog, and a variant of the waypoint symbology planned for the RAH-66. The simulation mission was built around a roughly 30 mile, low altitude navigation task in which the pilot had to overfly eight waypoints and return to the takeoff point. Twice during each mission, the pilot was asked to reconnoiter for parked tanks. The reconnaissance task was used with the situation awareness self-evaluations, and it increased the difficulty of the navigation task by making the pilot deviate from the course.

There were three alert display conditions. In the baseline condition no alerts were presented. In the full screen alert condition, all display symbology flashed. In the localized alert condition, only the alert symbol flashed. Following the alert, the pilot performed a procedural task, consisting of the entry of five, randomly chosen digits. Alerts were presented during four mission segments in order to evaluate the effect of the frequency of alert presentation on the quality of the data.

The alert could provide information about the task to be performed. A "partial information" condition informed the pilot whether digit entry would be performed left-to-right or right-to-left. A "no information" condition only alerted the pilot to the need to perform the digit entry task. The conditions are described in greater detail in the method section of the present report.

Evaluation of the display symbology was performed without reference to the workload and situation awareness metrics, and those results are presented elsewhere. We report objective performance measures in the present report only

when to compare them to the performance self-evaluations. We present an overview of the results here only to provide a context for the metrics evaluation. Both navigation displays improved navigation performance as compared to a baseline having only a compass and stationary map. The alert task took about 10 sec on the average.

# METHOD

## Apparatus

Helicopter Simulation - The investigation was conducted using the six-degree-of-freedom vertical motion simulator (VMS) with a rotorcraft cockpit. The VMS is unique among flight simulators in its large range of motion. This motion provides flight cues to the pilot.

A simulated rotorcraft cabin, the RCAB, was configured as a single-pilot cockpit with a four-window computer generated display, consisting of three forward view, CRT displays, spanning 27° X 147° and one CRT chin window on the right side (26° X 22°). The out-the-window imagery was generated using an Evans and Sutherland ESIG 3000 image generator.

The primary inputs to the motion base are the aircraft translational and rotational accelerations calculated by the math model for the pilot position. Appendix A contains a summary of the motion gains. The simulated aircraft was a UH-60A. Rotor, engine, and transmission sounds were simulated. Conventional helicopter controls were used. The stick-to-visual throughput time delay was approximately 72 msec.

Panel instruments were displayed on two 14 in diagonal color CRTs. The right CRT displayed generic, basic flight instruments (see Figure 1). The left CRT displayed a moving map of the visual data base (see Figure 2). The map showed major terrain features and major roads, shown in light blue. The planned course was shown in red. A compass rose was shown in the upper right hand corner. A gray square overlaid the high resolution area at the center of the data base. In the visual data base, this was a high definition rendering of a small village, that could not be represented on the map. A digital range indicator in the lower left corner indicated the size of the displayed area in nautical miles. When the alert task was presented, the required input and the pilot's response were overlaid on the bottom of the map. The helmet mounted display system consisted of the helmet, helmet display unit and head position sensing system of the AH-64 integrated helmet and display sight system.

Helmet Display Symbology - Helmet display symbology (see Figure 3) was based on the AH-64 pilot night vision system (PNVS), cruise mode symbology. This included a compressed 120° compass at the top of the display, digital torque and airspeed on the left, digital altitude and analog, "radar" altitude and vertical speed on the right. Altitude was above ground level for both digital and analog displays. A dashed line gave a rough indication of pitch and roll, referenced to the display frame. A diamond indicating the position of the aircraft's nose is the only head slaved PNVS symbology. The test formats had symbology added for the CDI, Waypoint, and alert displays.

Alert Task Symbology – There were five alert type conditions. A No-alert condition provided a flying performance baseline. Alert flash (Localized and Full-Screen) was crossed with alert information content (No-Info and Partial-Info) to yield four experimental conditions. Table 1 shows the design of the display test. While this experimental design is irrelevant to the evaluation of workload and situation awareness metrics, we present it to provide a context.

The alert was presented solely on the helmet mounted display, with a digit entry task to simulate the procedural response presented on the panel.

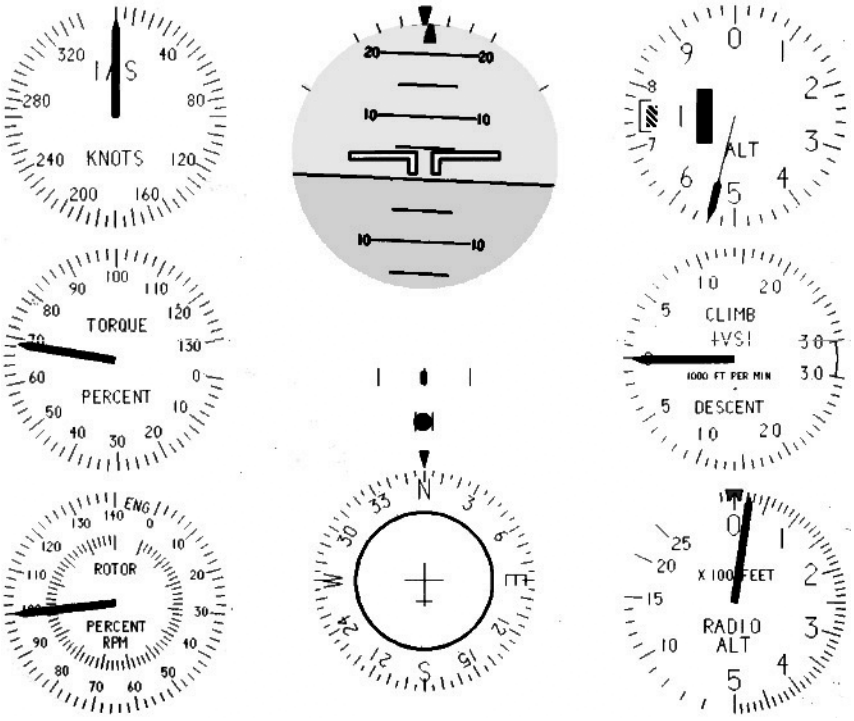



Figure 1 Simulated Instrument Panel. Instruments Consisted of White and Colored Graphics on a Black Background, Color Scheme has been Revised for Better Printing.

In all Alert Task conditions, a flashing letter was presented in the bottom center of the display, the position corresponding to the TADS field of view box (see Figure 4). The letter was approximately 3° high. In the Localized alert condition, this symbol constituted the entire alert. In the Full-Screen alert condition, all symbology flashed. Flashing alternated between the center and periphery of the display, that is, when the center brightness was high, the periphery was low, and vice versa. Central symbology consisted of the horizon line, nose diamond, alert letter, and the waypoint when it was present. All other symbology was peripheral. "High" brightness was the brightness set by the pilot. "Low" brightness voltage was 30% of high. This design ensured that some symbology would always be fairly bright and that no symbology would ever be completely off. Flash rate was set to be noticeable but not disturbing, at three Hz with a linear ramp up and down. Three alert symbols were used. In the No-Information condition, and upper case "N" indicated an alert. In the Part-Information condition, an "L" indicated an alert and that numbers for the "procedural" task were to be entered left-to-right, while and "R" indicated right-to-left entry.

Navigation Symbology – There were three navigation display conditions. A "Visual" navigation condition used the basic PNVS symbology without the

waypoint symbol. A "Waypoint" condition used a waypoint marker overlaid on the visual scene. The "CDI" condition incorporated symbols into the compass display indicating bearing to waypoints and course deviation. These two displays also included an arrival time clock in the upper right that showed the pilot's instantaneous arrival time error, up to +/-99 sec. Arrival time error was simply the difference between the target arrival time and the arrival time computed from current speed and distance remaining. In an actual mission speed would vary on each navigation leg, and so arrival time would be needed for each leg. In the simulation, planned speed was constant across legs, so only segment arrival time was displayed.

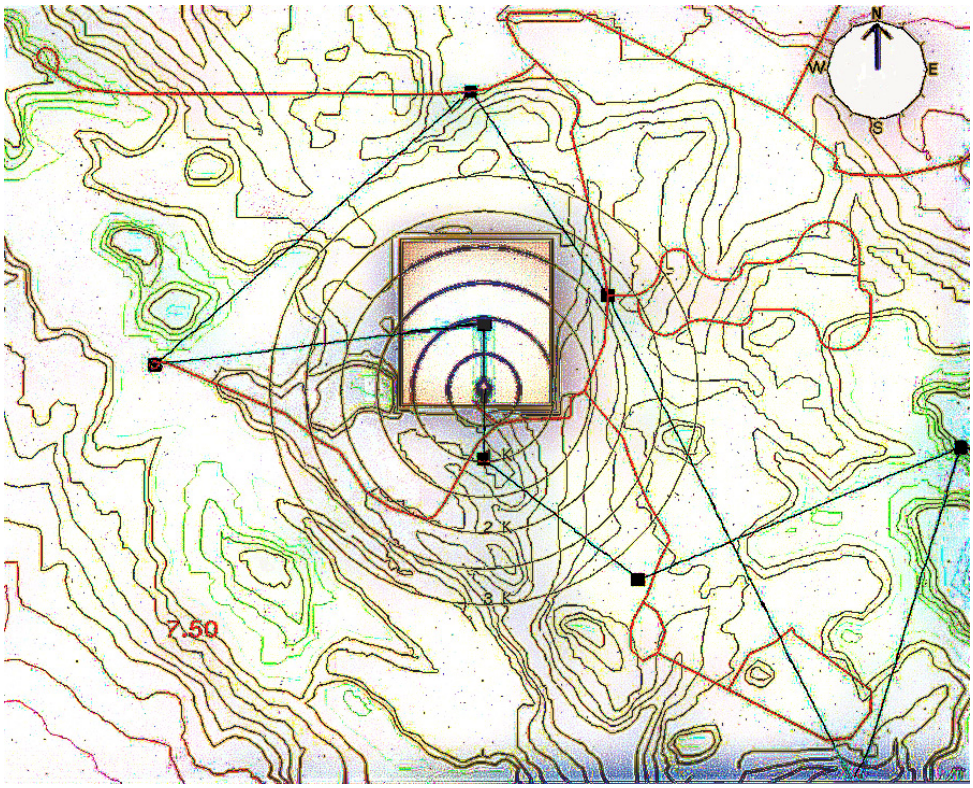



Figure 2. Map Display. Terrain Contours Consisted of Colored Lines on a Black Background. Roads Were Teal. Planned Route Was Red. Color Scheme has been Revised for Better Printing.

Waypoint Symbology - The Waypoint symbology consisted simply of a pennant displayed at the geographical location of each waypoint and the altitude of the aircraft. The pennants were maintained as moving models by the image generation system but were displayed by the Silicon Graphics computer that drove the helmet display. Each pennant was shaped like an arrow that pointed toward the next waypoint (see Figure 5). Because the pennants were maintained as part of the visual data base, all were displayed continuously, and their size decreased with range.

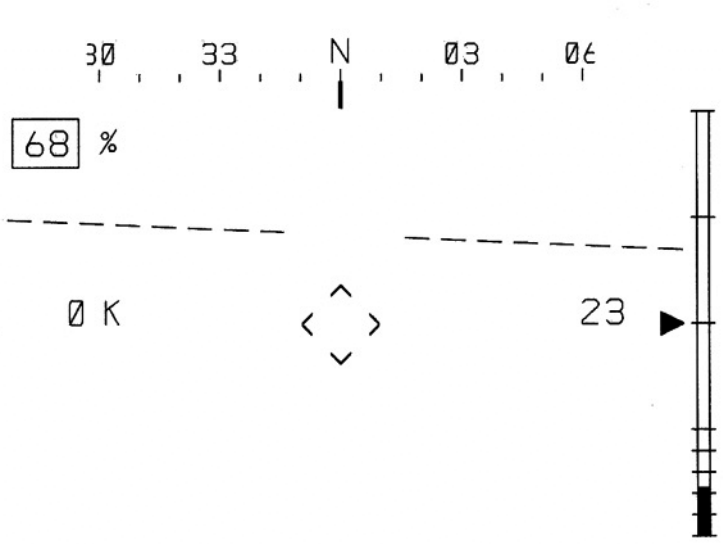



Figure 3. Basic Helmet Display Symbology.

CDI Symbology - The CDI symbology mimicked a conventional, panel mounted, course deviation indicator (CDI) (see Figure 6). A tail beneath the compass lubber line pivoted to point toward the planned course (shows no error in figur. A carat (**^**) indicated the heading to the current waypoint (as in the basic PNVS). A circle (**o**) indicated the heading to the next waypoint. Both waypoint symbols edge limited.

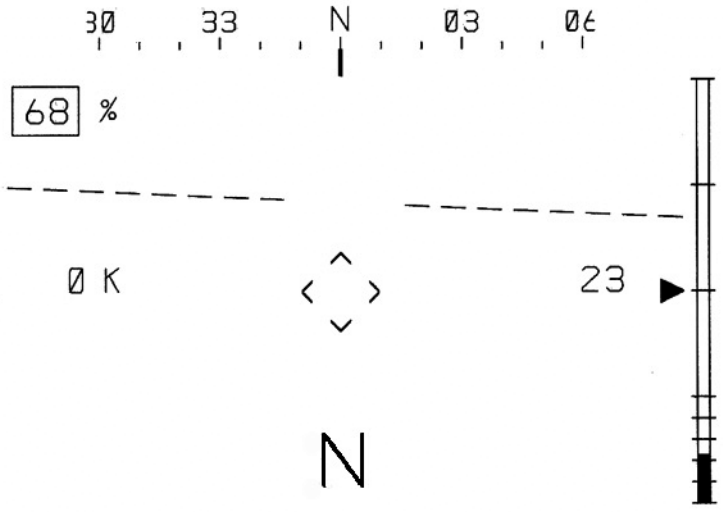



Figure 4. Helmet Display Alert Symbol

Alert Task Operator Interface – The pilot reacted to the alert symbol on the helmet mounted display by performing a task overlaid on the map display using a keypad. The keypad was located on a console next to the collective lever and contained a button to allow the pilot to acknowledge the alert along with a 10-key numeric pad. When the pilot acknowledged the alert, the helmet display

symbology ceased flashing, but the alert symbol remained present. Acknowledgement also caused the task display to appear, superimposed on the map display. The task display consisted of three lines of text. The top line was the word "LEFT" or "RIGHT," indicating the direction in which the pilot was to enter numbers using the 10-key pad. The second line showed five digits, selected randomly with replacement, which the pilot was to enter. The third line showed five blanks, corresponding to the five digits to be entered. As the pilot entered each correct digit, it was displayed in the correct blank. Incorrect entries were ignored. Once the pilot entered the fifth correct digit, both the map and helmet mounted displays returned to the nominal state.

Table 1. Alert Conditions.

| | No Alert | No Information | Partial Information |
|---|---|---|---|
| No Alert | **Baseline** | | |
| Localized Alert | | **X** | **X** |
| Full-Screen Alert | | **X** | **X** |

We developed this task in an attempt to capture the salient aspects of a procedural task. These include a structured sequence of actions, determination of required response, and making the required response. By giving the pilot a complex task, we hope to learn more than simply how fast the pilot could react to the alert. For this task the sequence was acknowledge alert, determine direction of digit input (left or right), input five digits. Unlike an actual flight

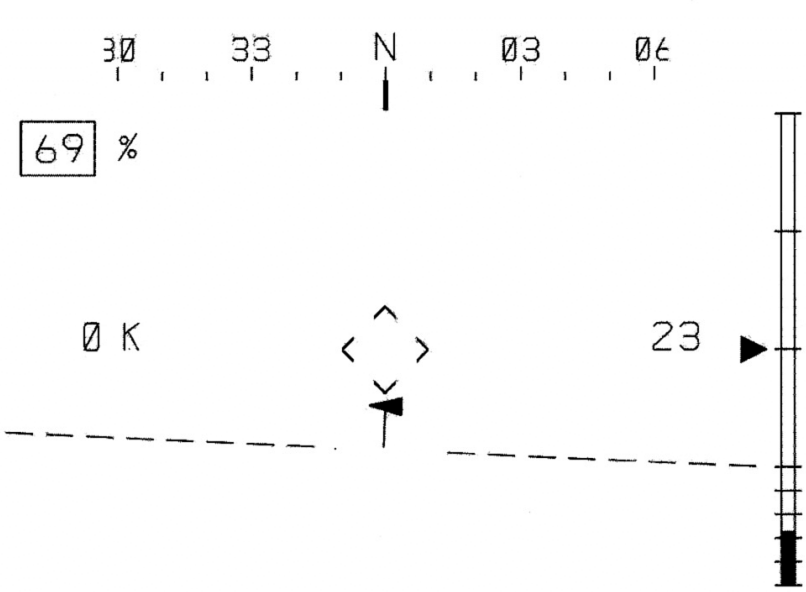



Figure 5. "Waypoint" Navigation Symbol.

procedure, this task had no mission performance consequence. The task could not be performed incorrectly, since wrong inputs would not be accepted; and slow performance had no effect on the simulation.

Experimental Tasks – A standardized flying task was developed, that consisted of segments. These were (1) take-off and cross-country navigation with reconnaissance for tanks, (2) low altitude track following, (3) cross-country flight with reconnaissance for tanks, (4) bob-up, and (5) return to and landing in the village. The simulator detected when the aircraft passed within 1000 ft of each waypoint and automatically advanced to the next segment. Data were collected on segments one through four.

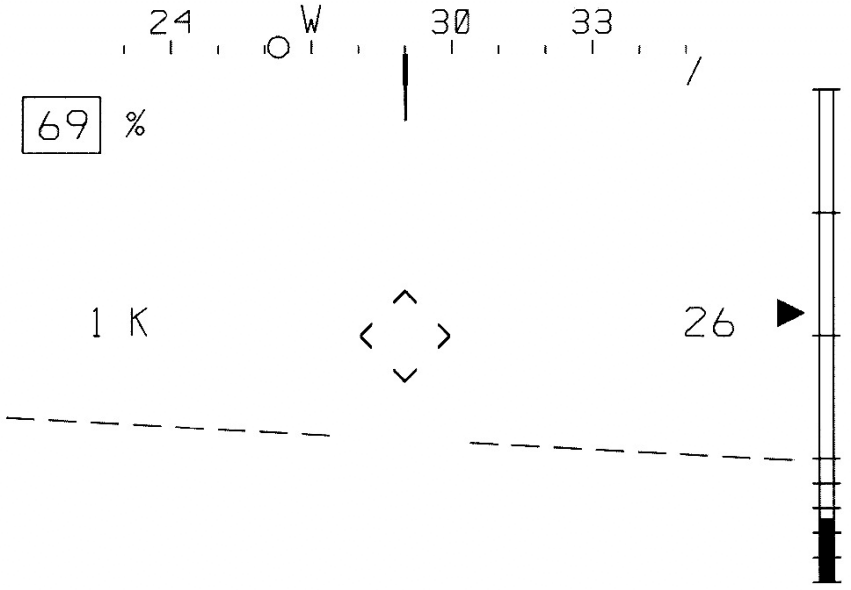



Figure 6. "CDI" Navigation Symbology. Current Waypoint Symbol is Edge Limited on Right and shows only one leg of the carat (^).

Twelve distinct missions were created by varying the waypoints and tank laydowns on the two cross-country navigation segments. The take-off, track following and bob-up segments were the same on each mission. The number of waypoints on each navigation segment was also constant, but their location was varied. The pilot took off and flew to the first waypoint, a beacon just outside the village. From there he turned to the first of two variably placed waypoints. From the second variable waypoint, the pilot flew to the start of the track. After executing the track following task, the pilot turned to the first of two variable waypoints on the second navigation segment, ending at the bob-up site. Following the bob-up, the pilot returned to the village by a constant route and landed.

Navigation and Reconnaissance Tasks – On each of the two cross-country segments the pilot flew from a constant location waypoint (the beacon or end of the track) to an ending point (the start of the track or the bob-up) by way of two

intermediate waypoints whose location varied from mission to mission. He was to reconnoiter for tanks, whose location and number varied. The reconnaissance task was intended to increase workload by providing additional tasking and to make the navigation task more challenging by forcing the pilot off the course. When he completed his reconnaissance, he depressed the microphone switch on the cyclic grip to mark the report time and reported the number of tanks seen, over an open microphone. He was given no feedback regarding his performance on any part of this task, save the arrival time clock on the test navigation displays. In preliminary runs with unlimited visibility and a moving map display, we found navigation to be very easy. Therefore, we reduced visibility to 5000 ft and rendered the map stationary. A single alert was presented during each navigation segment. The duration of each navigation segment was 10 to 15 minutes.

Track Following Task – In the track following task, the pilot was to follow the centerline of a two-lane road at nap-of-the-earth altitude over rolling terrain. At an assigned airspeed of 40 kt, this task took about six minutes. An alert was programmed to occur randomly during each consecutive one-minute interval. If the pilot flew too fast, later alerts would not occur.

Bob-up – The bob-up task was developed from the Aeronautical Design Standard-33 (ADS-33, 1994) bob-up task used in handling qualities evaluations. In our task the pilot was to hover 10 ft above the ground, to ascend rapidly to 50 ft above ground level, to hover for 10 sec, and to descend to the low hover. Out-the-window cues to altitude and position were provided by hover boards in front of the aircraft and walls off the right nose (see Figure 7). The alert task occurred three times in the bob-up, making the bob-up very different from the ADS-33 task. The visual and motor workload were very much higher due to the requirement for precision flying combined with the complicated alert response. The alerts were keyed to phases of the bob-up maneuver. One alert was programmed during the ascent, one during the high hover, and one during the descent. These phases were very short in duration, so the pace of activity was very rapid, and timing was critical. Total task duration was less than 30 seconds.

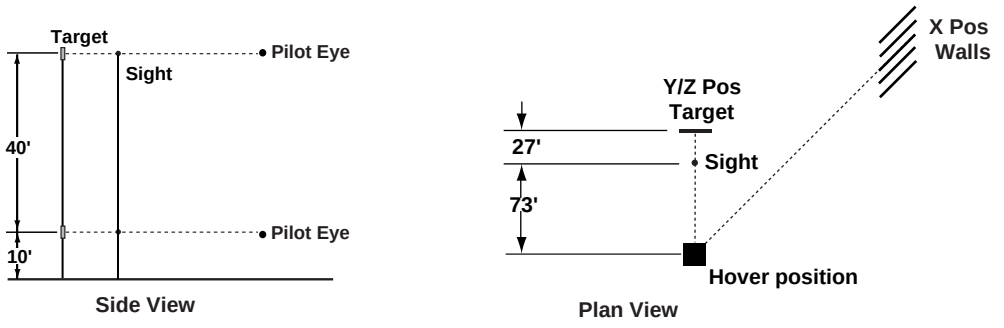



Figure 7. Hover Boards and Walls to Provide Position Cues for Bob-up Task.

Procedure

Pilots participated in pairs. Each pilot performed one to three missions and then took a break while the other pilot flew. The duration of each pair's simulation period was four days. Missions lasted about 30 minutes. Following each

mission the pilot gave his workload and performance self-evaluation ratings without receiving any feedback on his performance. Pilots received written instructions explaining the goal of the research and the tasks that they would perform (see Appendix B). They then performed familiarization flights until they were ready to begin practicing the experimental tasks.

Pilots received paper maps that duplicated the cockpit map in order to familiarize themselves with the mission before hand. They were also allowed to make a list of the waypoints for each mission to take into the cockpit. As the pilot passed each waypoint, he was to depress the microphone switch and state the waypoint name.

Three practice missions were provided. These missions had no tanks present, and the pilots did not perform reconnaissance. The pilots flew practice missions in each of the navigation display and alert conditions until they and the experimenter felt that they were ready to proceed to the experimental trials. Pilots gave no workload or self-evaluation ratings of the practice runs.

The order of presentation of the navigation display conditions for data collection was balanced by a Latin Square (n = 12). Presentation of alert conditions was balanced to control for first order effects. Pilots flew three to six experimental missions per day for a total of 12 experimental missions.

Data collected included workload ratings and performance self-evaluations, mission time of waypoint passage reports, mission time of reconnaissance reports, reconnaissance reports, and automatically recorded flight performance data. The pilots gave workload ratings and performance self-evaluations orally over the intercom, and the experimenter transcribed them into a log, at the end of each mission.

Data were also collected on responding to the alerts. These data are reported in a separate report (in preparation). Alert response performance was not the subject of workload and situation awareness ratings.

Workload ratings were based on a 101 point, interval scale. A rating of zero indicated no workload. A rating of 100 indicated maximum possible workload. We did not formally define zero and 100% workload but left them to the pilot's interpretation. Pilots rated their workload on six phases of the mission. Ratings were made for "Input," that is, gathering information, "Central," that is, thinking about the task, "Output," that is, making control inputs or other actions, and "Time" pressure.

The performance self-evaluations were based on defined error criteria for each task rated. Task performance ratings used three ordinal level values similar to those used in handling qualities evaluation. "Desired" meant the highest level of performance. "Adequate" meant performance that was acceptable but not of the highest level. "Outside of Acceptable" meant unacceptable performance. Error scores defining each performance level are presented in Table 2.

Table 2. Error Criteria for Performance Self-Evaluation Ratings.

| | | D (Desired) | A (Acceptable) | O (Outside of Acceptable) |
|---|---|---|---|---|
| Recon | Accuracy (% tanks detected) | >90% | 75% - 90% | <75% |
| | Timeliness (report time after first detection | <20 sec | 20 - 40 sec | >40 sec |
| Navigation | Accuracy (maximum deviation from course) | <100 ft | 100 - 200 ft | >200 ft |
| | Timeliness (at track and bob-up) | <+/- 10 sec of assigned time | <+/- 20 sec of assigned time | >+/- 20 sec of assigned time |
| Bob-up | Height | <+/- 3 ft | <+/- 6 ft | >+/- 6 ft |
| | Time | <+/- 4 sec | <+/- 6 sec | >+/- 6 sec |
| | Position | <+/- 6 ft | <+/- 10 ft | >+/- 10 ft |

# RESULTS

## Reconnaissance Situation Awareness

We meant the reconnaissance task to perform several functions in the simulation. It served as a workload enhancer in its own right, and it increased the difficulty of the navigation task by forcing the pilot to deviate from the planned course in search of the targets. It also provided a way of testing the performance self-evaluation situation awareness metric. In some sense reconnaissance performance is the most interesting application of the technique because the pilot has no feedback about the correctness of his reports. So must base his evaluation on his perception of the situation when he performed the task.

Following each simulated flight, the pilot provided a rating of the quality of his performance on the two reconnaissance tasks. We compared this rating with one made by the experimenter using the actual detection data. Scoring of the detection data was complicated by the fact that the pilots sometimes wanted to report individual tanks and sometimes wanted to report "platoons." We accommodated the pilots' desire to report platoons by devising a scheme for converting the number of tanks observed to number of platoons. A platoon consists of four tanks, so the pilots were told to consider four or fewer tanks to be one platoon. When a pilot found more than four tanks, he was to consider groups of four to be platoons and any remaining tanks to be elements of a platoon. In effect then the pilot divided the number of tanks by four and rounded any remainder up.

Table 3 shows the correct reconnaissance report, in tanks or platoons, for each mission laydown in the first column. The top entry of each laydown shows the first reconnaissance segment, and the bottom row shows the second segment. The entry shows 'number of tanks'/'number of platoons.' The pilot self-evaluation data is shown in the shaded part of Table 3. Number of vehicles reported is in the same format as that for the laydowns. The capital letter in the bottom row shows the pilot's evaluation of his performance. The lower case letter in the top row shows the experimenter's evaluation based on the actual report.

Objective criteria were needed in order to evaluate reconnaissance performance. Criteria were provided to the pilots for "Desired," "Adequate," and "Outside of adequate" performance. The criteria against which the pilots scored themselves had only addressed missed tanks, but in some cases, pilots reported too many tanks. Therefore the experimenter's criteria used percentage deviation from actual without regard to the sign of the deviation. So reporting one tank too many was the same as reporting one tank too few.

Reports of platoon and individual tanks were combined by multiplying the deviation from the correct number of platoons by four and treating the result as a deviation in the number of tanks reported. So one platoon was converted to four tanks. The last two rows of Table 3 show the proportion of correct detections and the proportion of correct evaluations across the twelve missions.

Table 3. Comparison of Pilots' Reconnaissance Performance Self-Evaluation with Actual Performance. Numbers indicate 'tanks'/'platoons.' Letters indicate performance evaluations. "---" Indicates Missing Data.

| Laydown # Pilot Response # Tanks / # Pltns | Pilot 1 Report T/P Rate | Pilot 2 Report T/P Rate | Pilot 3 Report T/P Rate | Pilot 4 Report T/P Rate | Pilot 5 Report T/P Rate | Pilot 6 Report T/P Rate | Performance Rating |
|---|---|---|---|---|---|---|---|
| 1 6 / 2 | / 1 o | / 2 d | 6 / d | 5 / o | 3 / o | / 2 d | Actual |
| 3 / 1 | 0 / A | / 1 D | 3 / D | 0 / O | 3 / D | / 1 D | Pilot |
| 2 4 / 1 | / 2 o | / 2 o | 4 / d | 4 / o | 0 / o | / 2 o | Actual |
| 9 / 3 | / 3 D | 0 / D | 9 / D | 4 / A | 4 / D | / 2 A | Pilot |
| 3 4 / 1 | / 2 o | / 2 o | 4 / d | 4 / o | 3 / a | / 1 o | Actual |
| 7 / 2 | / 3 D | / 2 D | --- D | 0 / A | 8 / --- | / 3 A | Pilot |
| 4 10 / 3 | / 4 o | / 4 a | 10 / d | 6 / o | 10 / d | / 4 a | Actual |
| 8 / 2 | / 3 D | / 2 D | 8 / D | 5 / D | 8 / D | / 2 A | Pilot |
| 5 11 / 3 | / 4 o | 4 / o | 10 / d | 8 / a | 8 / o | / 3 d | Actual |
| 7 / 2 | / 3 D | 11 / D | 7 / D | 7 / D | 0 / O | / 2 D | Pilot |
| 6 7 / 2 | / 3 o | 3 / o | 7 / d | 5 / a | 5 / o | 0 / o | Actual |
| 5 / 2 | 0 / D | 4 / D | 5 / A | 5 / D | --- O | / 1 A | Pilot |
| 7 11 / 3 | / 4 o | --- | 10 / d | 7 / o | 8 / a | / 4 a | Actual |
| 5 / 2 | 3 / D | --- | 6 / D | 3 / A | 4 / D | / 2 D | Pilot |
| 8 5 / 2 | / 2 d | 0 / o | 6 / d | 3 / o | 0 / o | / 2 d | Actual |
| 5 / 2 | / 2 D | 0 / O | 5 / D | 2 / O | 2 / D | / 2 A | Pilot |
| 9 5 / 2 | / 2 o | 3 / o | 4 / d | 2 / o | 4 / o | / 2 o | Actual |
| 11 / 3 | 4 / D | 0 / --- | / 3 D | 9 / D | 0 / O | / 1 A | Pilot |
| 10 11 /3 | 3 / o | 8 / o | 9 / a | 5 / o | 7 / a | / 2 a | Actual |
| 5 / 2 | 2 / D | 3 / D | 4 / D | 4 / D | 5 / D | / 2 D | Pilot |
| 11 4 / 1 | / 2 o | 3 / o | 3 / a | 3 / a | 5 / d | 0 / o | Actual |
| 10 / 3 | / 3 D | 7 / D | 0 / D | 9 / D | 10 / D | / 2 A | Pilot |
| 12 8 / 2 | / 3 o | 9 / d | 5 / a | 5 / o | 6 / a | 5 / o | Actual |
| 6 / 2 | / 2 D | 6 / D | 6 / D | 5 / D | 6 / D | 16 / D | Pilot |
| Proportion Correct Detections | 0.38 | 0.54 | 0.86 | 0.66 | 0.64 | 0.71 | |
| Proportion Correct Self Evaluations | 0.08 | 0.30 | 0.67 | 0.17 | 0.42 | 0.25 | |

The pilots showed a substantial range of performance in detection of the tanks. The worst pilot detected fewer than 40% of the targets, while the best detected up to more than 80%. The pilots' self-evaluations showed an even greater range, and they were highly correlated with performance ($r = 0.78$, $p < 0.07$ (Beyer, 1966); see Figure 8). This result may actually support the conclusion that pilots who were less effective in the reconnaissance task were in fact less situationally aware. That is, they were less able to tell when they had performed a thorough search, while the more effective pilots were able to tell how thoroughly they had searched, even without the benefit of feedback on their effectiveness. The performance self-evaluation does appear to be a good measure of situation awareness.

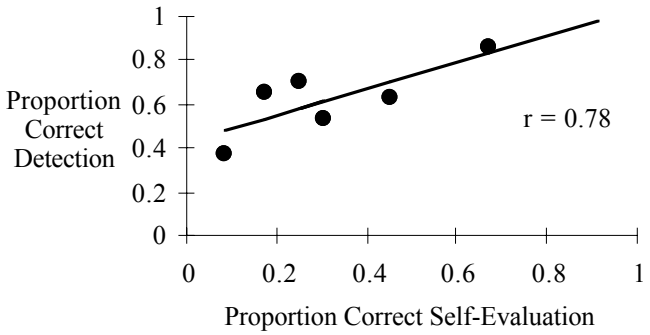



Figure 8. Relationship between Self -Evaluation and Actual Performance. Correlation is Significant at $p < 0.07$ (Beyer, op cit).

Bob-Up Situation Awareness

We encountered numerous technical difficulties with the bob-up task. These stemmed largely from the difficulty of responding to the alerts while maintaining aircraft control. The alert task required the pilots to remove their hand from the collective, which had a severe impact on aircraft control. The time required to perform the alert response task was so long that the response to an alert in one phase of the bob-up (e.g., ascent) was not completed until a subsequent phase (e.g., high hover). This caused problems with the timing of data collection. Nevertheless, we were able to get self-evaluation data which shed some light on how the pilots reacted to an overwhelming task ensemble. We shall examine three aspects of bob-up performance: time in the high hover, maximum altitude error in the high hover, and maximum horizontal position error.

Examining task performance, we see that the pilots found this task ensemble nearly impossible (see Table 4). They met the time criteria for holding the high hover on fewer than 50% of the trials. They met altitude criteria on only about 15% of the trials, and none succeeded in meeting the horizontal position criteria. Performance this poor raises concern about a flaw in the simulation, but we checked the flight model and hardware thoroughly and found none. When we examined the self-evaluations, the reason for the poor performance became apparent. The pilots simply neglected the precision hover task (presumably to perform the alert response). This task had been included to examine the validity of very frequent alerts (which would provide a larger amount of data). In the event, the task was too unrealistic. Normally the pilot would simply have landed and performed the alert task on the ground.

The pilots were only moderately accurate in the self-evaluations of time in hover. Their self-evaluations were correct on only about 62% of the trials. Their accuracy was particularly low for performance in the "Adequate" category. They showed the best accuracy when they rated their performance "Outside of Acceptable." Station keeping errors of up to several hundred percent of the

desired let them achieve 92% accuracy. This compares with accuracy on the reconnaissance task of 71% and 29% for "D" and "O" respectively.

Time was the factor to which the pilots attended most, as indicated by self-evaluation accuracy. Overall self-evaluation accuracy was low even for horizontal position, where all performance was "O," and errors were as large as 100 times the desired criterion. This was true despite the fact that on the bob-up performance is easily judged looking out the window. By contrast pilots had to infer performance on the reconnaissance task, since they had no way to know how many targets they had failed to see. Yet the self-evaluation of reconnaissance performance was more strongly related to actual performance, and the pilots were much better able to judge good performance in that task.

Table 4. Performance and Pilot Self-Evaluation on the Bob-up Task. Proportions Sum to less than 1.0 because of Missing Data.

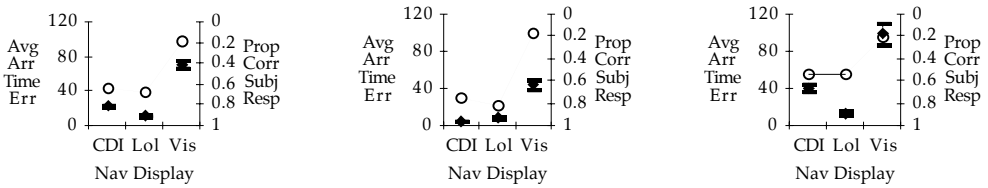
| Performance Factor | Proportion of D / A / O Trials | Proportion of D / A / O Self-Evaluations | Proportion of D / A / O (Overall) Correct Self-Evaluations |
|---|---|---|---|
| Time in High Hover | 0.45/0.17/0.38 | 0.49/0.26/0.25 | 0.65/0.29/0.92 (0.62) |
| Max Z Error in High Hover | 0.15/0.26/0.55 | 0.34/0.45/0.23 | 0.28/0.26/1.0 (0.50) |
| Max Horizontal Position Error | 0.00/0.00/0.95 | 0.23/0.42/0.35 | 0.00/0.00/1.00 (0.35) |

We infer from these results that the pilots simply neglected the aircraft control portion of the bob-up task ensemble. This is an unusual finding, which may reflect their priorities in performing the task and the unrealistic nature of the task. As a result they failed to detect even very large flying performance errors. The pattern of self-evaluation deficiency supports this conclusion. Time, which can be perceived with little attention, was the most accurately evaluated factor. Altitude error, which was strongly cued by the ground and the hover boards in front of the aircraft, was judged somewhat less well. Horizontal position required the pilot to attend to the forward hover boards and the walls located well off the right nose. This required both head movement and mental effort. Apparently, the pilots were unable to integrate this activity with their other tasks. While this bob-up task left much to be desired for evaluating alerting display formats, the self-evaluations were effective in allowing us to measure the pilots' situation awareness and to relate that to the patterns of performance.

Navigation Situation Awareness

The navigation task was the only task on which the pilot's situation awareness and self-evaluation were subject to direct manipulation. This was done through the design and content of the display format on the HMD. Two aspects of the

display format could affect self-evaluation and situation awareness. These were the navigation symbology (i.e., the CDI or Waypoint) which told the pilot where he was relative to the planned course, and the arrival time clock. The former allowed him to monitor and control his course-following performance, while the latter simply allowed him to monitor his arrival time error. Both display formats did improve performance significantly over a visual navigation condition in which the pilot had only a static map (see Figure 9).




a. Segments 1 & 2 b. Segment 1 c. Segment 2

Open Symbols Indicate Proportion Correct Self-Evaluations; scale is inverted. Filled Symbols Indicate Mean Arrival Time Error and 95% Confidence Interval of Mean.

Figure 9. Performance Self-Evaluations and Actual Arrival Time Performance on Two Navigation Segments Combined (a) and on Segments 1 (b) and 2 (c) Individually.

The HMD navigation display formats also had a pronounced effect on the pilot's self-evaluations. The navigation display formats improved the accuracy of the pilot's self-evaluations ($F(2, 27) = 14.31$ $p < 0.05$, SAS, 1992). Self-evaluations using the two navigation display formats were significantly different from those in the visual navigation condition by a Student Neuman-Keuls test, but were not different from each other. Based on inspection of the data in Figure 9, we expected possible effects of navigation segment and of display format by segment interaction. The segment effect failed to reach significance, however ($F1, 27) = 3.1$, $p > 0.08$), since there was no effect of segment on the Waypoint self-evaluations. The data were insufficient to test for interactions.

The segment effect might have shed light on the question of why the navigation display formats supported more accurate self-evaluations. The navigation display formats provided better information by indicating deviation from the planned course. They also provided performance data in the form of the arrival time clock. In the bob-up task, we have seen that the presentation of easily apprehended performance information can improve self-evaluation accuracy and that poor self-evaluations can reflect a failure to note presented information. In the reconnaissance tasks, we have seen that pilots can make accurate self-evaluations based on their situation awareness, in the absence of any performance feedback. Either of these options is possible in the case of the navigation display formats.

The second reconnaissance task was much more difficult than the first, leading to worse performance. This increased reconnaissance difficulty led to poorer arrival time performance because the pilots spent more time at reconnaissance and strayed further from the course. As they spent more time on reconnaissance, the pilots became disoriented and less situationally aware, which might have

been reflected in less accurate self-evaluations. So even though the performance feedback was equal on both segments, the self-evaluations would have been less accurate on the more difficult second segment. The trend in the data supports this option more than the notion that pilots simply read their performance from the arrival time clock, although the data were statistically marginal.

Workload Measurement

Workload Assessment was not a primary objective of the research. It was included to determine the utility of the DRAWS workload scale. The set of flight tasks used in the research was well suited to this determination in that it included tasks covering a large range of expected workload demand.

The bob-up is a very high workload task, requiring the pilot to perform precise control of aircraft position in x, y, and z and to maintain yaw angle. The task requires the pilot to attend to out-the-window cues that are widely separated in space and that vary continuously. As we implemented the bob-up task ensemble, workload demand was greatly increased by the need to respond to frequent alerts. These alerts forced the pilot both to look into the cockpit and to remove the left hand from the collective.

Track following was another continuous control task, but the precision required was considerably less, and the attentional demand was lower. In addition the workload imposed by the alerts was lessened by the lower frequency with which they occurred.

The reconnaissance task was embedded in a larger navigation task. The workload rating covered both the reconnaissance task and its interaction with the navigation task. The workload demand of this task ensemble was probably the one we understood least well going into the research. While workload demand of navigation and cross country flight was fairly low in the simulated environment, the reconnaissance task could impose significant workload depending on how well targets were hidden, and time pressure on the reconnaissance task could increase as the pilot fell further behind on the navigation task. The alert task was not a significant workload inducer, with only one response being required over the 10 to 15 minute navigation segment.

The take-off and landing tasks had fairly low workload, with no traffic, communication procedural tasking. The alert task was not presented on either take-off or landing segment.

The first question to answer is whether DRAWS is sensitive to the workload differences described above. We examined this question by performing a separate analysis of variance (Pilot X Task) for each workload factor in DRAWS (SAS, 1992). Figure 10 shows the mean ratings on each workload factor for each flight task. All workload factors were highly sensitive to variation in workload demand across the flight tasks. Within each workload factor, tasks in Figure 10 labeled with the differing letter are significantly different from each other according to a post-hoc Student Newman-Keuls test ($p < 0.05$). For all workload factors, the bob-up produced the highest ratings, and the take-off produced the lowest ratings, as expected. For the Central, Input, and Output factors, the two reconnaissance segments, the track and the landing produced intermediate

ratings. The Time factor was unique in that the track and landing tasks produced low ratings on this factor.

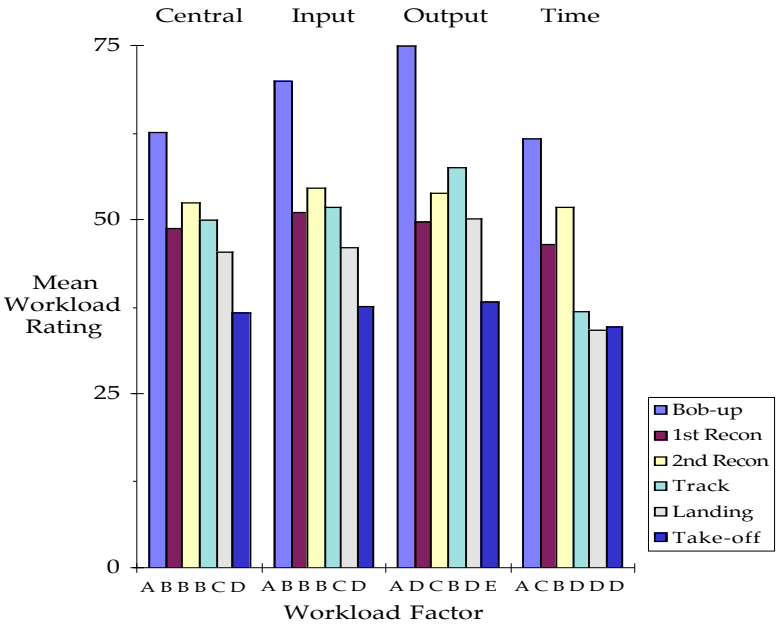



Figure 10. Mean DRAWS Workload Ratings for the Six Flight Tasks. Tasks within each Factor having the Same Letter Identifier are not Significantly Different from each Other According to a Post-hoc Student Newman-Keuls Test ($p < 0.05$, SAS, 1992).

The similarity of the ratings for the four workload factors led us to question the independence of the four factors. The pattern of ratings is nearly identical for the Central and Input factors, and the pattern differs only slightly from them for the Output factor. We examined this relationship by correlating the ratings given by each pilot to each task on the four factors (see Table 5). We examined the reliability of these correlations by calculating a mean correlation across pilots and testing the difference of this mean from zero using a simple, paired, Student's t-test ($p < 0.05$, see Table 6, Hays, 1963). The Central, Input, and Output factors were all significantly correlated with each other. The Time factor was not significantly correlated with any of the other factors.

An unusual aspect of DRAWS is its 101 point range of possible ratings on each factor. Most rating scales use a range closer to Miller's (Psych Rev, 1956, 63, 81-97) magical number seven. For example, the Task Load Index (TLX, Hart and Staveland, 1988) uses 21 points (zero to 100 in increments of five), and the Subjective Workload Assessment Technique (SWAT, Reid et al, 1981) uses three. The zero and 100 points of the DRAWS scale are considered to indicate no demand and 100% demand respectively, although it is not intuitively clear what constitutes "no demand" or "100% demand." This question notwithstanding, we can ask how the 101 point scale affects the performance of DRAWS as a measurement tool.

Table 5. Individual Pilot Correlations between DRAWS Workload Factors.

| | | | |
|---|---|---|---|
| Input | 0.97 | | |
| | 0.92 | | |
| | 0.91 | | |
| | 0.97 | | |
| | 0.89 | | |
| | 0.89 | | |
| Output | 0.86 | 0.94 | |
| | 0.92 | 0.93 | |
| | 0.86 | 0.91 | |
| | 0.82 | 0.88 | |
| | 0.73 | 0.93 | |
| | 0.32 | 0.57 | |
| Time | 0.62 | 0.58 | 0.49 |
| | 0.18 | 0.01 | -0.10 |
| | 0.69 | 0.68 | 0.89 |
| | -0.10 | -0.11 | -0.35 |
| | 0.04 | -0.05 | 0.14 |
| | 0.32 | 0.63 | 0.20 |
| | Central | Input | Output |

We created artificial "n-point" scales by rounding the ratings, where n varied from two points to 101 points. This kept the scales comparable in the magnitude of the workload scores they produced but varied their resolution. The result is shown in Figure 11. Except for some small and inconsistent artifacts of the rounding, reducing the number of rating points did not change the performance of DRAWS until the number of points was reduced to two per factor. The effect was similar for all four workload factors, so we collapsed across factors in Figure 12. Here we see that the two-point scale yields artificially inflated workload estimates compared to the other scales.

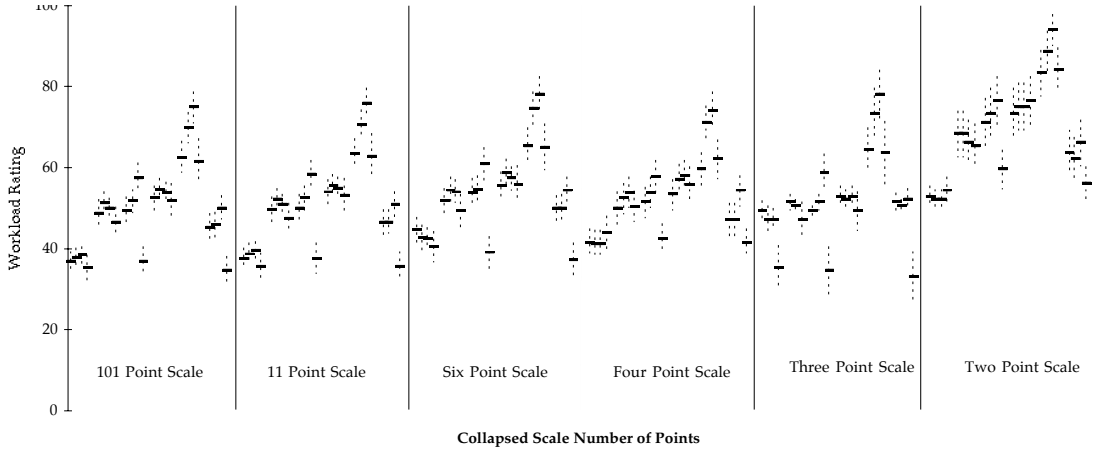



Figure 11. Effect of Decreasing Scale Resolution on DRAWS Sensitivity. Within each Scale Group, Tasks Are Clustered in Order of Occurrence in the Flight: T/O, Recon 1, Trk, Recon 2, B-U, Lnd. Within Each Task, Workload Ratings Are Input, Central, Output, Time. Mean Ratings and 95% Confidence Intervals of the Mean Are Shown.

The DRAWS ratings reflected the workload that we had expected from the flight tasks. The pilots were able to use DRAWS to give reasonable workload estimates with little training. Forgetting over the course of the flight did not have a negative impact on the sensitivity of the ratings. It is interesting to note that two of the scales, Output and Time, reflected differences between the two reconnaissance segments. In fact the pilots commented that the second segment was more difficult than the first, and the arrival time error data showed poorer performance on the second segment. Based on pilot comments and experimenter observations we believe that the increased workload stems from the targets being more difficult to find on the second segment. This caused the pilots to spend more time searching and to wander further from the course. The increased Time rating therefore appears reasonable, while the smaller increase in the Output rating is harder to interpret.

Three of the four DRAWS scale are highly correlated with each other, and this correlation is consistent across pilots. These scales could likely be collapsed into one "Performance Workload" scale with little loss of information. A potential benefit of collapsing the scales would be a reduction in the likelihood of obtaining spurious significant differences. Combining redundant scales would allow adding other workload factors without increasing the workload impact of DRAWS, itself.

The unusually large DRAWS scale provides no clear benefit over smaller scales. There is a risk to the larger scale that cannot be demonstrated by collapsing the scale as we have done. That risk arises because the pilots do not use all, or even most, of the DRAWS scale points. Since each pilot selects the points he will use, each in effect uses a unique scale of his own design. This can only increase the variance in the ratings and possibly obscure workload. Figure 12 illustrates how pilots' style affects the variability of ratings. The range of workload ratings given by a pilot is the difference between the maximum rating and the minimum

rating. The range of ratings given by the six pilots varied from 50 points to 80 points. They varied to a greater extent in their tendency to use ratings divisible by five, but not by 10. Less than 1% of pilot 1's ratings were divisible by only five, while over 57% of pilot 3's ratings were divisible by only five. We group the pilots by their tendency to use ratings divisible by five (pilots 3 and 6) or not (pilots 1, 2, 4, and 5) and determined an effective number of points in each pilot's scale by dividing by five or 10. The effective number of points was higher for pilots that used ratings divisible by five. If only ratings divisible by 10 were allowed the variation between pilots was greatly reduced.

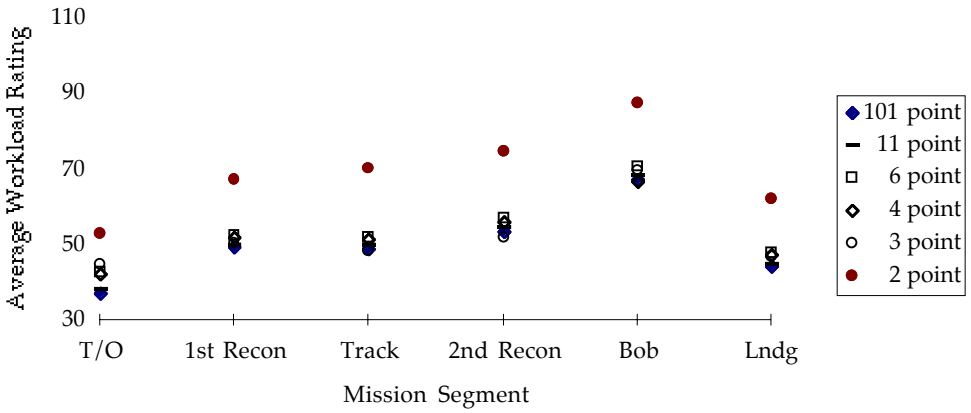



Figure 12. Effect of Scale Resolution on DRAWS Sensitivity Collapsed across Factors. Means of Four Factor Means Are Shown.

Table 6. Average Correlations between DRAWS Workload Factors. Means of Six Individual Pilot Correlations Are Shown. Student's t (paired) Values for the Average Correlation Are Shown in ().

| Workload Factor | Mean r (t(5)) | Mean r (t(5)) | Mean r (t(5)) |
|---|---|---|---|
| Input | 0.92 * (53.4) | - | - |
| Output | 0.75 * (7.6) | 0.86 * (13.3) | - |
| Time | 0.29 (2.1) | 0.29 (1.7) | 0.21 (1.1) |
| * t(5) > 2.228, p <0.05 | Central | Input | Output |

The DRAWS ratings provided reliable workload measures, that are in accord with our expectations of the tasks. The pilots were able to use the DRAWS with little practice. Therefore DRAWS can be considered a usable workload measurement tool. It is not, however, without its issues. Three of its four scales are highly correlated with each other. Based on this correlation, we feel that the Input, Output, and Central scales all tap into a single, more global, task performance workload factor. Raters unable to distinguish finer details within this factor, at least along the dimensions established by DRAWS. There would seem to be little benefit gained from gathering data on redundant scales.

Further, use of multiple, redundant scales poses the risk that chance, erroneous workload effects might be observed.

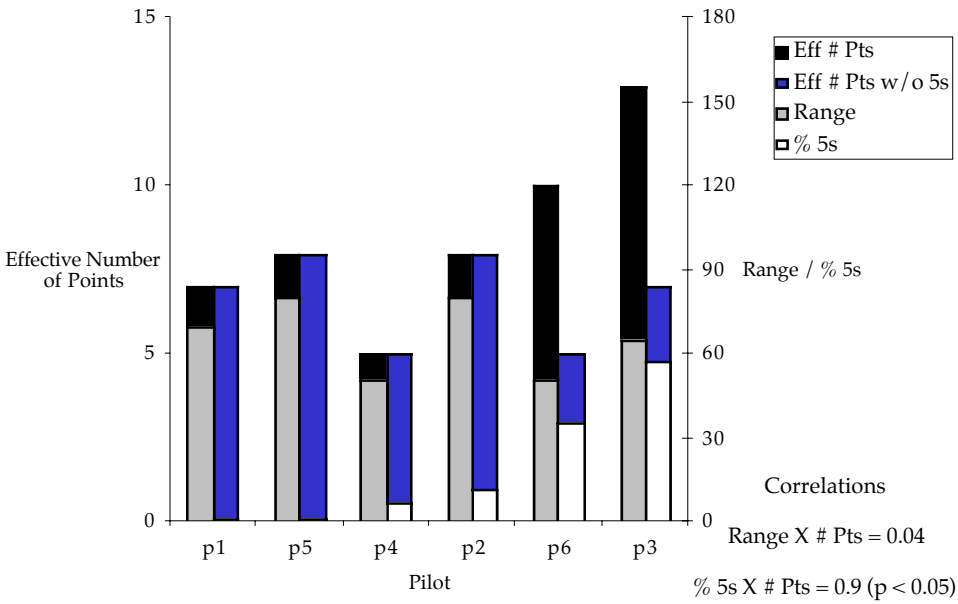



Figure 13. Effect of Pilot Rating Style on Variability of Workload Ratings.

Another issue is the large range of each rating scale, 101 points. This range is so large that raters use only a small portion of the possible ratings, usually only those evenly divisible by 10, but occasionally those evenly divisible by five. Since each rater can in effect design his own scale, there is a risk that real workload effects might be masked by increased interrater variability. The present research finds no benefit of increased resolution to justify the risk.

## DISCUSSION

Performance self-evaluations offer an appealing approach to the measurement of situation awareness. They are easily gathered at the end of task performance, and they require no special training on the part of the operators. The question is whether self-evaluations truly reflect situation awareness or simply the operator's ability to note and remember the required information. The answer from this research is that performance self-evaluations do reflect operator's situation awareness. The interpretation of self-evaluations, however, can be complicated.

We saw that in the case of the reconnaissance self-evaluation, pilots were able to make self-evaluations that were highly correlated with actual performance. The validity of the self-evaluation as a measure of situation awareness is shown by the fact that the pilots could accurately infer performance quality without direct feedback. Its utility lies in showing that poorer reconnaissance performance was linked to a lesser ability on the part of the pilot to evaluate the thoroughness of the search.

In the case of the bob-up, we were confronted with a task ensemble that the pilots found nearly impossible to perform. In this case the self-evaluations were not strongly related to actual performance, but there was a strong indication from the self-evaluations that the poor performance arose from the pilots' simply not attending to the flying task.

The self-evaluations allowed us to examine the contribution of novel information displays to pilots' situation awareness in the navigation task. The novel displays presented course deviation and arrival time performance information. The situation awareness measure distinguished the novel displays from a baseline display, but it was insensitive to differences between novel displays.

When we examined the effect of difficulty of an embedded reconnaissance task on navigation performance and self-evaluation, we saw that both poorer performance and poorer situation awareness resulted from the more difficult reconnaissance task. There was an indication that the pilots' self-evaluations were based more on the information in the navigation display than on the arrival time clock. Also the information presentation in the "Waypoint" display appeared to be degraded less by reconnaissance task workload than did that in the "CDI".

We easily used the DRAWS techniques for measuring pilot workload, and the pilots found them easy to use. DRAWS provided workload rating that conformed generally to our expectations of the tasks rated. So DRAWS is a usable system, but does it provide any benefit over existing workload measurement tools.

DRAWS is comparable to other workload measurement tools, both in ease of use and in ease of data reduction. It differs from other measurement tools in its emphasis on the information processing aspect of task performance while ignoring more "human" aspects such as "Frustration," used in TLX and "Psychological Stress," used in SWAT. These factors have proven to have utility

in understanding workload; so there elimination can be viewed as a liability. By contrast DRAWS focuses very heavily on the process of moving from perception to action, with three separate factors, "Input," "Central," and "Output." We found these three factors to be highly correlated with each other. We interpret these correlations to reflect the fact that operators have difficulty introspecting about these components of the performance process. On balance it would seem that the decomposition of workload factors in DRAWS offers little if any benefit over existing workload measurement tools and may result in the loss of some important workload information.

# CONCLUSIONS

The present research examined the effectiveness of a novel workload measurement scale , the Defense Research Agency Workload Scale (DRAWS) and a novel approach to evaluation of situation awareness, the performance self-evaluation. Both techniques were evaluated in a simulated cross-country navigation mission with embedded tasks intended to increase pilot workload.

We examined the performance self-evaluation metric for situation awareness in three very different tasks, reconnaissance, precision bob-up, and cross-country navigation. In each of these tasks, the self-evaluation proved to be a simple and effective measure of situation awareness when analyzed with relevant performance data.

In the reconnaissance task, the self-evaluation was highly correlated with the accuracy of pilots' reports, even though the pilots had no means of determining their report accuracy. We concluded that the self-evaluation was sensitive to the pilot's perception of the local area and his awareness of how thoroughly he had searched, a true measure of situation awareness.

In the bob-up task, a lack of correlation between the self-evaluation and station keeping performance was interpreted to suggest an explanation for this poor performance. The bob-up task was a severe overload situation, and the pilots simply ignored cues to hover performance in order to concentrate the procedural task.

In the navigation task, the self-evaluation allowed us to separate the effects of two displays intended to aid performance. Changes in the accuracy of self-evaluations when the difficulty of a secondary, reconnaissance task increased showed that situation awareness information provided by navigation symbology was more important than performance feedback provided by an arrival time clock.

In all three instances that we examined, we were able to make inferences about pilots' situation awareness by comparing the self-evaluations to actual performance. The inferences varied with the type of situation, the type of performance. and the sources of variation. In all cases the technique allowed data collection with minimal impact on the task situation or performance since self-evaluations were collected following performance.

The DRAWS scale was found to be easily understood by the pilots. They were able to provide workload ratings that showed reliable differences between the experimental tasks. Three of the DRAWS scales, "Input," "Output," and "Central" were highly correlated with each other and may tap into a single, overarching workload factor associate with the mechanics of task performance.

DRAWS functions effectively as a two-factor workload rating scale. Expanding the scale to include three highly correlated factors may have increased the likelihood of obtaining spurious high workload ratings.

DRAWS uses a 101-point rating scale, intended to provide intuitive zero percent and 100 percent demand anchors. While the intuitive value of these anchors was not assessed, the pilot's ignored most of the rating options consistently. In effect

each pilot designed his own scale by the selection of a small subset of rating points. This may increase between rater variance and lessen the effectiveness of DRAWS.

REFERENCES


Aeronautical Design Standard 33D, Handling Qualities Requirements for Military Rotorcraft, U. S. Army Aviation and Troop Command, St. Louis, MO, 1994.

Beyer, W. H. (ed.) Chemical Rubber Co. Handbook of Tables for Probability and Statistics, The Chemical Rubber Co., Cleveland, 1966.

Endsley, M. R. "Toward a Theory of Situation Awareness in Dynamic Systems," Human Factors, 1995a, 37(1), 32 – 64.

Endsley, M. R. "Measurement of Situation Awareness in Dynamic Systems," Human Factors, 1995b, 37(1), 65 – 84.

Eubanks, J. L. and Killeen, P. R. "An Application of Signal Detection Theory to Air Combat Training," Human Factors, 1983, 25, 449 – 456,.

Fracker, M. L. Measures of Situation Awareness: an Experimental Evaluation, AL-TR-191-0127, Armstrong Laboratory, Wright-Patterson AFB, Ohio, 1991.

Hart, S. G. and Staveland, L. E. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in Hancock, P. A. and Meshkati, N. (Ed.) Human Mental Workload, North Holland Press, Amsterdam, 1988.

Hays, W. L. Statistics, Holt, Rinehart, and Winston, New York, 1963.

Long, G. M. and Waag, W. L. "Limitations on the Practical Applicability of d' and β Measures," 1981, Human Factors, 23, 285 – 290.

McMillan, G. R., Bushman, J., and Judge. C. L. A. "Evaluating Pilot Situational Awareness in an Operational Environment," Situation Awareness: Limitations and Enhancement in the Aviation Environment, 1996, AGARD-CP-575 (K1-1 – K1-6), Nueilly-Sur-Seine, France.

Miller, G. A. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," Psychological Review, 1956, 63, 81 – 97.

Reid, G. B., Shingledecker, C. A., and Eggemeier, T. F. "Application of Conjoint Measurement to Workload Scale Development," Proceedings of the Human Factors Society 25th Annual Meeting, Rochester, NY, 1981.

SAS/STAT User's Guide, Version 6 (2), SAS Institute Inc, Cary, NC, 1992.

# APPENDIX A

# MOTION GAINS

*** SLOW GAINS

GXS=.4
GYS=.8
GZS=.5
GPS=.05
GQS=.05
GRS=.5

*** SLOW CORNER FREQUENCIES

OMEGXS=1.5
OMEGYS=.6
OMEGZS=.3
OMEGPS=.5
OMEGQS=.7
OMEGRS=.5

*** FAST GAINS

GXF=.4
GYF=.6
GZF=.9
GPF=.5
GQF=.8
GRF=.4

*** FAST CORNER FREQUENCIES

OMEGXF=1.5
OMEGYF=.6
OMEGZF=1.4
OMEGPF=.85
OMEGQF=.85
OMEGRF=.7

*** DAMPING RATIO

ZETAX=.707
ZETAY=.707
ZETAZ=.707
ZETAP=.707
ZETAQ=.707
ZETAR=.707

*** SLOW AND FAST "SPEEDS" FOR INTERPOLATION OF SLOW AND FAST GAINS

VGFAST=10.
VGSLOW=0.

*** SLOW AND FAST "SPEEDS" FOR INTERPOLATION OF SLOW AND FAST CORNER FREQUENCIES

VWFAST=10.
VGWLOW=0.

*** TC FORWARD PATH GAINS

GXTC=1.
GYTC=1.

*** TC FEEDBACK GAINS

GKTCFB=.1

*** RESIDUAL TILT CORNER FREQUENCIES. DAMPING RATIO, AND GAINS

*** CORNER FREQ. OF LARGE AXIS RT LOWPASS FILTER

OMEGLRT=2.

*** CORNER FREQ. OF SMALL AXIS RT LOWPASS FILTER

OMEGSRT=2.

*** DAMPING RATIO OF LARGE AXIS RT LOWPASS FILTER

ZETAR1=.707
ZETAR2=.707

*** ROLL AND PITCH RT GAINS

GPRT=1.
GQRT=1.

GKTCFB=0.5

# Appendix B

## Pilot Instructions

# Introduction and Instructions to Experimental Pilots

This experiment is the start of a program of research, conducted jointly by the Army and NASA, to develop principles for the presentation of symbolic information on a helmet mounted display. The current work uses a production AH-64 IHADSS, but the goal is to extend the work with an advanced color, wide field of view, binocular display system. The research is conducted on NASA's Vertical Motion Simulator, the largest motion based simulation in the world.

The research will examine presentation of two types of information on the helmet display, alert information and navigation information. Alert information will be presented in two modes. In a one mode a flashing letter presented at the bottom of the HMD will indicated that you should come into the cockpit to perform a procedural task. In the second mode the flashing letter will be accompanied by flashing of all HMD symbology. Two navigation display conditions will be used. In one waypoint data will be presented by markers on the HMD that have been located and scaled to fit into the visual scene. In the second an "instrument" located at the top of the HMD will present waypoint information. The test conditions for both experimental questions will be embedded in a simulated cross-country flight mission of about 20 minutes duration.

## Cross-Country Flight

Prior to each flight, you will receive a map showing the route you are to fly. You will maintain an altitude of 100 ft agl and an airspeed of 80 kt. At each waypoint you will make a radio call to inform us that you have reached the waypoint. You should make a list of the waypoints and headings to aid you in the cockpit. Along the route you will perform two experimental flight tasks, that are not part of the cross-country flight. These tasks are a track following tasks and a bob-up. Your flight time from the take-off to the track and from the track to the bob-up position will be specified for each mission. You will also be asked to reconnoiter two areas of interest along each route and report back the number of tank platoons in each area. Performance specifications are given in the table at the end of this document.

## Track Following Task

You will follow a track on the ground maintaining a comfortable and an altitude of 100 ft AGL.

## Bob-up Task

You will establish a stable hover at 10 ft AGL in position in front of the bob-up tree. At this point make your radio call to tell us that you have reached the waypoint. This call will initiate the task. You will then climb to and altitude of 50 ft AGL and hover for 10 sec. Mark the start and end of the hover and the end of the task by pressing the xmit switch. Performance specifications are given in the table at the end of this document.

Procedural Tasks and Alert Displays

We have developed a laboratory task which is intended to demand your attention as would an in-flight procedure. This task will be presented on the left panel CRT, which normally displays the moving map. The task requires you to enter five digits on the numeric keypad located on the side console. The word "Left" or "Right" displayed above the number indicates whether they are to be entered from left to right or from right to left. The system will only respond to correct entries, which will be displayed on the panel. After you have entered all five digits, the map display will reappear.

An HMD alert will signal when you are to perform this task. This signal will consist of a flashing letter at the bottom of the HMD. In on condition, only this letter will flash. In a second condition, all HMD symbology will flash. The symbol may provide some information about the task to be performed. In one condition, the letter, "L" or "R" will indicate left-to-right or right-to-left entry. In a second condition, the letter "N" will indicate only that the number entry task is to be performed. Prior to beginning the number entry task, you must "clear" the alert by pushing the "Cancel" button next to the numeric pad.

The sequence of events for this task is as follows:

1. HMD flashes and number entry task replaces moving map
2. Pilot "clears" alert by pushing "Clear" button
3. Pilot enters five numbers on keypad
4. Panel display reverts to moving map.

Navigation Displays

There are two HMD navigation displays.

A "waypoint" display indicates the location of the current waypoint by a symbol floating above its location. The symbol will point left or right depending on the direction of turn required after passing the waypoint. A timer will track your on-time performance. A time of arrival error counter located in the upper right portion of the HMD tells how many seconds early (+) or late (-) you are.

A "CDI" display shows your deviation from the route, heading to the current waypoint, heading to the next waypoint, and on-time status. Deviation is indicated by a pointer located below the lubber line. This pointer indicates the direction to the course. Maximum scaled deviation is 2000 ft when the pointer is 60 degrees from vertical. The current waypoint is indicated by a carat on the compass scale. The next waypoint is indicated by a circle on the compass scale. These symbols edge limit if the waypoint is off scale. On-time status is indicated as on the "Waypoint" display.

Performance Criteria and Workload

You will be asked to evaluate your performance against the criteria in the table below. You will be asked to rate your workload on six phases of the mission. Your ratings will be made on a scale from 1 to 100. Ratings will be made for "Input," that is, gathering information, "Central," that is, thinking about the task, "Output," that is, making control inputs or other actions, and "Time" pressure.

Performance Criteria.

| | | D (Desired) | A (Acceptable) | O (Outside Acceptable) |
|---|---|---|---|---|
| Recon | Accuracy (% tanks detected) | >90% | 75% - 90% | <75% |
| | Timeliness (report time after first detection | <20 sec | 20 - 40 sec | >40 sec |
| Navigation | Accuracy (maximum deviation from course) | <100 ft | 100 - 200 ft | >200 ft |
| | Timeliness (at track and bob-up) | +/- 10 sec of assigned time | +/- 20 sec of assigned time | >+/- 20 sec of assigned time |
| Bob-up | Height | +/- 3 ft | +/- 6 ft | >+/- 6 ft |
| | Time | +/- 4 sec | +/- 6 sec | >+/- 6 sec |
| | Position | +/- 6 ft | +/- 10 ft | >+/- 10 ft |

# APPENDIX C

# ANALYSIS OF VARIANCE TABLES

General Linear Models Procedure

Dependent Variable: Central

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 35 | 67582.91 | 1930.94 | 21.70 | 0.0001 |
| Error | 367 | 32659.37 | 88.99 | | |
| Corrected Total | 402 | 100242.29 | | | |

| R-Square | C.V. | Root MSE | C Mean |
|---|---|---|---|
| 0.67 | 19.18 | 9.43 | 49.19 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **PILOT** | **5** | **32259.59** | **6451.92** | **72.50** | **0.0001** |
| **TASK** | **5** | **25377.85** | **5075.57** | **57.04** | **0.0001** |
| **PILOT*TASK** | **25** | **10930.24** | **437.21** | **4.91** | **0.0001** |

General Linear Models Procedure

Dependent Variable: Input

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 35 | 75261.09 | 2150.32 | 28.35 | 0.0001 |
| Error | 367 | 27837.82 | 75.85 | | |
| Corrected Total | 402 | 103098.90 | | | |

| R-Square | C.V. | Root MSE | C Mean |
|---|---|---|---|
| 0.73 | 16.81 | 8.71 | 51.81 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **PILOT** | **5** | **22736.81** | **4547.36** | **59.95** | **0.0001** |
| **TASK** | **5** | **39703.39** | **7940.68** | **104.69** | **0.0001** |
| **PILOT*TASK** | **25** | **14888.403** | **595.54** | **7.85** | **0.0001** |

General Linear Models Procedure

Dependent Variable: Output

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 35 | 80687.79 | 2305.65 | 26.49 | 0.0001 |
| Error | 367 | 31939.57 | 87.029 | | |
| Corrected Total | 402 | 112627.36 | | | |

| R-Square | C.V. | Root MSE | C Mean |
|---|---|---|---|
| 0.72 | 17.26 | 9.33 | 54.04 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **PILOT** | **5** | **12494.33** | **2498.87** | **28.71** | **0.0001** |
| **TASK** | **5** | **50870.45** | **10174.09** | **116.90** | **0.0001** |
| **PILOT*TASK** | **25** | **19959.04** | **798.36** | **9.17** | **0.0001** |

General Linear Models Procedure

Dependent Variable: Time

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 35 | 100338.58 | 2866.82 | 18.82 | 0.0001 |
| Error | 367 | 55910.79 | 152.35 | | |
| Corrected Total | 402 | 156249.38 | | | |

| R-Square | C.V. | Root MSE | C Mean |
|---|---|---|---|
| 0.64 | 27.77 | 12.34 | 44.44 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **PILOT** | **5** | **28253.46** | **5650.69** | **37.09** | **0.0001** |
| **TASK** | **5** | **44329.14** | **8865.83** | **58.20** | **0.0001** |
| **PILOT*TASK** | **25** | **32718.48** | **1308.74** | **8.59** | **0.0001** |

General Linear Models Procedure

Dependent Variable: TIME

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 35 | 116145.66 | 3318.44 | 2.49 | 0.0002 |
| Error | 96 | 127728 22 | 1330.50 | | |
| Corrected Total | 131 | 243873.94 | | | |

| R -Square | C.V. | Root MSE | TIME Mean |
|---|---|---|---|
| 0.48 | -197.18 | 36.48 | -18.50 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **PILOT** | **5** | **18635.16** | **3727.03** | **2.80** | **0.0209** |
| **SEGMENT** | **I** | **21278.6** | **27278.67** | **20.50** | **0.0001** |
| PILOT*SEGMENT | 5 | 5590.99 | 1118.20 | 0.84 | 0.5243 |
| **NAV_DISP** | **2** | **25446.25** | **12723.13** | **9.56** | **0.0002** |
| PIL0T*NAV_DISP | 10 | 11614.43 | 1161.44 | 0.87 | 0.5611 |
| SEGMENT*NAV_DISP | 2 | 8158.94 | 4079.47 | 3.07 | 0.0512 |
| PIL0T*SEG*NAV_DIS | 10 | 15911.79 | 1591.17 | 1.20 | 0.3034 |

General Linear Models Procedure

Dependent Variable: Correct Self-Evaluation

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 35 | 13.74 | 0.39 | 1.98 | 0.0049 |
| Error | 93 | 18.42 | 0.20 | | |
| Corrected Total | 128 | 32.15 | | | |

| R -Square | C.V. | Root MSE | Corr Mean |
|---|---|---|---|
| 0.43 | 84.42 | 0.45 | 0.53 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| PILOT | 5 | 0.94 | 0.19 | 0.96 | 0.4486 |
| SEGMENT | 1 | 0.91 | 0.91 | 4.57 | 0.0351 |
| PILOT*SEGMENT | 5 | 1.60 | 0.32 | 1.61 | 0.1644 |
| **NAV_DISP** | **2** | **6.50** | **3.25** | **16.40** | **0.0001** |
| PIL0T*NAV_DISP | 10 | 1.16 | 0.12 | 0.59 | 0.8213 |
| SEGMENT*NAV_DISP | 2 | 0.79 | 0.39 | 1.98 | 0.1435 |
| PIL0T*SEG*NAV_DIS | 10 | 2.05 | 0.21 | 1.04 | 0.4184 |

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



**1. AGENCY USE ONLY** *(Leave blank)*

**2. REPORT DATE**
October 1999

**3. REPORT TYPE AND DATES COVERED**
Technical Memorandum

**4. TITLE AND SUBTITLE**
Situation Awareness and Workload Measures for SAFOR

**5. FUNDING NUMBERS**
581-31-22

**6. AUTHOR(S)**
Joe De Maio and Sandra G. Hart

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Aeroflightdynamics Directorate, U.S. Army Aviation and Missile Command, Ames Research Center, Moffett Field, CA 94035-1000

**8. PERFORMING ORGANIZATION REPORT NUMBER**
A-99V0037

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
National Aeronautics and Space Administration
Washington, DC 20546-0001 and U.S. Army Aviation and Missile Command, Redstone Arsenal, AL 35898-5020

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
NASA/TM–1999-208795
AFDD/TR-00-A-002

**11. SUPPLEMENTARY NOTES**
Point of Contact: Joe De Maio, Ames Research Center, MS 210-5, Moffett Field, CA 94035-1000
(650) 604-6974

**12a. DISTRIBUTION/AVAILABILITY STATEMENT**
Unclassified — Unlimited
Subject Category 53 Distribution: Standard
Availability: NASA CASI (301) 621-0390

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** *(Maximum 200 words)*

The present research was conducted in support of the NASA Safe All-Weather Flight Operations for Rotorcraft (SAFOR) program. The purpose of the work was to investigate the utility of two measurement tools developed by the British Defense Evaluation Research Agency. These tools were a subjective workload assessment scale, the DRA Workload Scale (DRAWS), and a situation awareness measurement tool in which the crews self-evaluation of performance is compared against actual performance. These two measurement tools were evaluated in the context of a test of an innovative approach to alerting the crew by way of a helmet mounted display. The DRAWS was found to be usable, but it offered no advantages over extant scales, and it had only limited resolution. The performance self-evaluation metric of situation awareness was found to be highly effective.

**14. SUBJECT TERMS**
Workload, Situation awareness, SA, HMD

**15. NUMBER OF PAGES**
46

**16. PRICE CODE**
A03

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | | |