Modified Cooper Harper scales for assessing unmanned vehicle displays

Citation: Birsen Donmez, M. L. Cummings, Hudson D. Graham, and Amy S. Brzezinski. 2010. Modified Cooper Harper scales for assessing unmanned vehicle displays. In Proceedings of the 10th Performance Metrics for Intelligent Systems Workshop (PerMIS '10). ACM, New York, NY, USA, 235-242.

As Published: http://dx.doi.org/10.1145/2377576.2377620

Publisher: Association for Computing Machinery (ACM)

Persistent URL: http://hdl.handle.net/1721.1/81763

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Modified Cooper Harper Scales for Assessing Unmanned Vehicle Displays

Birsen Donmez
University of Toronto
Dept. of Mech. and Ind. Engineering
Toronto, ON, Canada
1 (416) 978 7399
donmez@mie.utoronto.ca

M. L. Cummings
MIT
Dept. of Aero-Astro
Cambridge, MA, USA
1 (617) 252 1512
missyc@mit.edu

Amy S. Brzezinski
NASA Johnson Space Center
Expedition Vehicle Division
Houston, TX, USA
1 (281) 244 5780
amy.s.brzezinski@nasa.gov

Hudson D. Graham
MIT
Dept. of Aero-Astro
Cambridge, MA, USA
1 (617) 258 5046
hgraham@mit.edu

ABSTRACT
Unmanned vehicle (UV) displays are often the only information link between operators and vehicles, so their design is critical to mission success. However, there is currently no standardized methodology for operators to subjectively assess a display's support of mission tasks. This paper proposes a subjective UV display evaluation tool: the Modified Cooper-Harper for Unmanned Vehicle Displays (MCH-UVD). The MCH-UVD is adapted from the Cooper-Harper aircraft handling scale by shifting focus to support of operator information processing. An experiment was conducted to evaluate and refine the MCH-UVD, as well as assess the need for mission-specific versus general versions. Participants (86%) thought that MCH-UVD helped them identify display deficiencies, and 32% said that they could not have identified the deficiencies without the tool. No major additional benefits were observed with mission-specific versions over the general scale.

Categories and Subject Descriptors
C.4 [**Performance of Systems**]: Measurement Techniques
H.1.2 [**User/Machine Systems**]: Human Factors
H.5.2 [**User Interfaces**]: Evaluation/methodology

General Terms
Measurement, Performance, Experimentation, Human Factors, Standardization.

Keywords
Unmanned Vehicles, Display Design, Subjective Assessment, Rating Scale.

PerMIS'10, September 28-30, 2010, Baltimore, MD, USA.

1. INTRODUCTION
Most of today's unmanned vehicle (UV) systems require supervision from human operators, who use displays to monitor and command UVs. Because UV operations are and will be conducted with operators physically separated from the UVs, interfaces often do and will provide the only link between the operators and the UVs. It is imperative that these interfaces effectively support the UV operator's higher level cognitive processes.

Given the rapid development and deployment of UV systems [1], a relatively simple but diagnostic display evaluation tool can provide substantial benefits to the design and implementation of UV interfaces. Such a tool should help diagnose potential interface issues, address specific deficiencies, and suggest potential remedies to improve the interfaces early in the design process, for relatively low cost and effort.

Although a general scale could provide a universal way to assess UV displays and is an attractive option due to simplicity in both administration and analysis, the increasing diversity and specificity of UV missions may require specific scales for UV display assessment. Thus, one question that needs to be addressed in the development of a standardized UV display assessment tool is whether a tool tailored to a specific mission or vehicle provides more useful and accurate results. By using a UV/mission specific scale, developers may be able to pinpoint what UV and mission-specific elements the display is not supporting. However, such an approach could be more costly in terms of time and resources.

This paper presents a subjective evaluation tool, called the Modified Cooper-Harper for Unmanned Vehicle Displays (MCH-UVD), which aims to identify and diagnose specific areas where UV displays might not support operator cognitive processes. The MCH-UVD scale is designed to be presented to UV operators immediately after display use in order to quickly gather operators' judgments on how well an interface supports UV supervision and the overall mission. The aviation industry has long relied on a standardized measurement tool, called the Cooper-Harper scale [2], for subjective evaluation of aircraft handling characteristics by pilots. The Cooper-Harper scale provides a standardized

method to compare handling qualities across aircraft and help test pilots articulate specific aircraft handling problems. The MCH-UVD was modeled after the Cooper-Harper scale by adopting its flowchart format, but the focus was changed from manual aircraft handling to higher-level cognitive processes such as information analysis and decision making.

The paper also presents a preliminary experiment conducted to evaluate and refine the MCH-UVD, as well as to determine if creating a mission-specific MCH-UVD is necessary, or if the general form of the MCH-UVD is sufficient for different UV display evaluation.

2. MODIFIED COOPER-HARPER SCALE FOR UNMANNED VEHICLE DISPLAYS

An initial MCH-UVD adaptation was conducted by Cummings et al. [3]. Ten ratings were separated into four distinct blocks, three of which addressed a different stage of the human information processing model [4] (acquisition, analysis, and decision making), and one which represented acceptable display designs. Mandatory redesign was suggested for deficiencies in information acquisition or perception, and deficiencies in decision-making warranted improvement. The 10 categories were selected after a comprehensive usability literature review [5, 6], detailed UAV accident analyses, and extensive testing [3, 7].

This initial scale had some shortcomings. First, the scale used technical human factors terminology, with which UV operators are generally not familiar. Also, the scale did not address the action stage of the human information processing model [4], thus neglecting the effectiveness of display affordances in supporting operator tasks, which is a major consideration affecting control of UVs. Moreover, the information processing stages which were addressed by the scale were not completely divided between their representative three blocks, thus causing some overlap. Further, by prioritizing the scale strictly along the information processing model, some validity of the display deficiency severity was lost. Critical deficiencies threatening mission success can occur in any of the processing stages.

The general MCH-UVD diagnosis tool shown in Figure 1 represents a major redesign of the initial MCH-UVD. This redesign includes the use of more common language for descriptions. Like the Cooper-Harper scale that rated aircraft controllability on a scale of severity, the intent was to scale severity to reflect display support for safe mission completion. At the same time, the concept of addressing the human information processing model was maintained, as this is a critical component to UV display designs. The redesign also includes a specific UVD deficiency defined for each rating.

The diamond block questions on the left of the scale ask an operator 1) if the mission can be completed safely, 2) if the UV display aids in mission completion, and for applicable cases, 3) whether it aids in mission re-planning. Based upon the operator's answer; the tool guides the user to another question querying about display support of the mission. Within the individual diamond groups, the human information processing model is applied on a severity scale, from information acquisition to information analysis to decision-making and finally action-taking tasks. The deficiencies are deemed to be more severe for tasks earlier in the human information processing model, because if the display is flawed in supporting an earlier stage, it does not matter how good the display supports the later phase because the operator could be acting upon a flawed input.

When operators are directed to the right of the diamond block questions, they examine a set of descriptions pertaining to potential issues. Within each diamond block, operators choose between two to four different descriptions, each of which corresponds to a MCH-UVD rating.

2.1 Cannot Complete Mission Task Safely

2.1.1 Flawed Information Retrieval – Rating 10

A display is considered to be flawed in information retrieval when it is missing critical information, essential information cannot be located, or information becomes outdated because of long retrieval times. A display that requires extensive multi-layered search could receive a rating of 10 if searching results in long retrieval times. Generally, under this diagnosis displays do not provide operators with the necessary information they need for tasks, making higher-level information processing virtually impossible and increasing the likelihood of UV mission failure.

For example, among the major causes of the Three Mile Island disaster was the lack of display indications of the coolant level in the core and a relief valve failure [8]. The operators did not realize that the core was experiencing loss of coolant and implemented a series of incorrect actions that led to catastrophic consequences.

2.1.2 Action Confusion – Rating 9

UV displays that do not provide straightforward or intuitive ways to act upon decisions are classified as having confusing action implementation. These displays may have display affordances [9] that are difficult to find or use, or that are easy to incorrectly use, thus making operator tasks hard to perform or easy to erroneously execute. Mode confusion, which occurs when operators do not understand or confuse the mode they are in [10], is another possible outcome of this rating. Incorrect task performance because of poor affordances could threaten mission success, even when information acquisition, analysis, and decision-making have been efficiently and properly performed.

For example, in 2006 a MQ-9 Predator B UAV impacted the terrain in Arizona during a nighttime border patrol mission and destroyed the aircraft. Mode confusion was identified as one of the major factors [11]. For this flight, there were two nearly physically identical consoles for the pilot and the payload operator, with the ability to transfer control from one console to the other. The functionality differed vastly depending on the mode of operation. The throttle quadrant on the pilot's console provided UAV airspeed control, but for the payload operator's console, this same quadrant controlled the camera. Before transferring control from one console to the other, the condition lever position on the payload operator's console had to be matched to the current positioning on the pilot's console. When the pilot and payload operator swapped seats, the pilot forgot to match the controls, which led to mode confusion, and ultimately, the loss of the UAV.

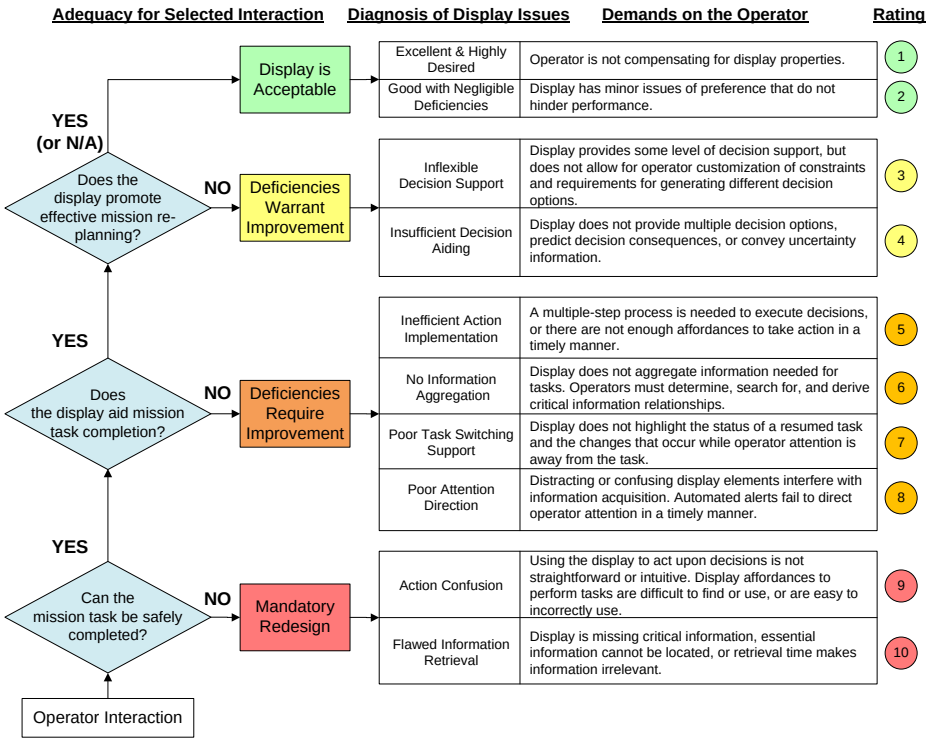

Figure 1. General MCH-UVD diagnosis tool.

2.2 Display does not Aid Mission Task Completion

2.2.1 Poor attention direction – Rating 8

These displays provide operators with the information they need, but contain distracting display elements or clutter, which could interfere with efficient information acquisition. Additionally, if the displays use automated alerting, these alerts do not attract operator attention in a timely manner. Under rating 8, information acquisition is hindered, but is still possible.

An illustrative case is the 1995 cruise ship Royal Majesty accident. After about 35 hours in transit from Bermuda to Boston, the ship grounded on the Nantucket shoals [12]. Shortly after departure, the GPS switched to dead-reckoning mode as it was no longer receiving satellite signals. None of the crew members were aware that GPS was in dead reckoning for the duration of the trip. There was a very small visual indicator on the GPS representing dead reckoning operations, but it was cluttered by other visual information. Moreover, the aural alarm indicating the switch to dead reckoning sounded only for a brief time in the chart room, but not on the bridge's central console where the watch officer stood. Thus, neither the visual display nor the auditory alert directed crew attention to the change in the system state.

2.2.2 Poor Task Switching Support – Rating 7

When the operator intentionally or unintentionally moves attention to one task to another, switching to the new task and switching back to the previous task comes with a cost [13-15]. UV displays receiving a rating of 7 have issues with supporting task switching. Although these displays, in part, support information analysis tasks, they do not clearly highlight the status of the resumed task and any changes that may have occurred while the operator's attention shifted. Display issues with task switching support could cause operators to make decisions and take actions based upon incorrect information, and may increase the time operators spend on analysis tasks. A classic human factors example of this problem is checklist interruption. A number of aircraft accidents have occurred due to such interruptions, particularly in the takeoff configuration when pilots are interrupted in the takeoff checklist process and forget to lower the flaps, resulting in stall and crash of the aircraft [16].

2.2.3 No Information Aggregation – Rating 6

UV displays that do not amass task information collectively or require operators to determine, search for, and derive critical information relationships are considered deficient in information aggregation. These displays do not suggest what information to analyze and do not co-locate different pieces of information related to a specific task, which is critical for effective decision making in dynamic environments [17]. Thus, the cognitive load of the operators increases as they have to determine what information to analyze, where the information is, and how to analyze it. These problems are exacerbated under time-pressure and dynamic settings, inherent characteristics of UV domains. For example, the poorly designed status display of Apollo 13, which lacked data history and a reference frame for oxygen tank pressure, did not aid the controllers to detect the imminent failure of a tank, leading to an emergency recovery of the spacecraft [18].

2.2.4 Inefficient Action Implementation – Rating 5

These displays either require unnecessary multiple step processes to execute actions, or do not provide enough affordances to take action. In these cases, displays generally support operator actions, but not efficiently, which could have negative effects on a UV

mission, particularly under time-critical situations. An illustrative example for this rating is the early version of BMW iDrive®, where drivers had to interact with a joystick and navigate through several screens for simple tasks such as changing the interior temperature.

2.3 Display does not Promote Effective Mission Re-planning

Re-planning decision support is not necessary in all UV missions. The N/A option allows the evaluator to bypass the question in cases where this functionality is not required such as in the case of small, handheld UAVs that are providing local imagery. UV decision support tools are generally supplemental tools, which could enhance the operator's interaction with the display [19, 20].

2.3.1 Insufficient Decision Aiding – Rating 4

These displays do not provide multiple decision options or predict the potential consequences of decisions. Additionally, these displays do not convey uncertainty information about decision alternatives, their potential consequences or about decision-making in general. Insufficient decision aiding might increase the likelihood of error, and potentially jeopardize mission success.

2.3.2 Inflexible Decision Support – Rating 3

Operators who rate a UV display as a 3 believe that it provides some level of decision-making support, but the display does not allow for operator customization of constraints and requirements to narrow down decision options. This inflexible decision support, as subjectively deemed by operators, is useful to the decision-making process but may not help operators make optimal decisions, which could potentially negatively affect a UV mission, particularly in time-critical situations.

2.4 Acceptable Displays

Acceptable displays include two ratings: good displays with negligible deficiencies (rating 2), excellent and highly desired displays (rating 1). A UV display receives a rating of 1 when the operator is not compensating for any deficient display properties. UV displays that receive a rating of 2 support human information processing through all four stages, but have very minor issues of preference that do not hinder operator performance. Example issues include preference of font style or size, display colors, or display element arrangement or sizing.

3. EVALUATING THE MODIFIED COOPER-HARPER SCALE

A human subject experiment was conducted to evaluate both general and UV/mission specific scales using two different UV displays. In addition to comparing general and specific scales, this experiment helped us better define the ratings. The MCH-UVD, presented in its general or specific form, was administered to participants as a post-test survey for evaluating two types of unmanned vehicles displays: UAV or UGV.

Figure 2 illustrates how the general MCH-UVD was modified to represent the specific unmanned ground vehicle (UGV) search mission utilized in the experiment. Defining the mission is a critical step for generating a specific MCH-UVD as the specific scale addresses the particular aspects of the UV/mission explicitly. Some of these aspects include the mission tasks that need to be safely completed by the UV, the critical information needed for the mission, the tasks/actions operators should perform, the system elements that require operator attention, the tasks that may require re-planning, and the uncertainties in decision-making.

3.1 Participants

Sixty participants completed the study. The participants consisted of 24 female and 36 male MIT students, ages ranging from 18 to 45 years (mean = 20.7, SD = 4.01). The experiment lasted 1 to 1.5 hours. The participants were compensated $10/hour and were eligible to win a $50 gift card based on their performance.

3.2 Apparatus

The UAV condition utilized the displays created for the Onboard Planning System for UAVs in Support of Expeditionary Reconnaissance and Surveillance (OPS-USERS) simulator, developed by Aurora Flight Sciences. OPS-USERS simulates UVs conducting search and tracking operations. The UGV condition was carried out on Urban Search and Rescue Simulator (USARSim), an open source, high fidelity simulation of urban search and rescue robots and environments [21].

The OPS-USERS display was presented on a 17-inch desktop screen, in 1600 x 1024 pixels and 32-bit resolution. The display for USARSim [22] was presented on a 17-inch desktop screen, in 1280 x 1024 pixels and 32-bit resolution. Participants controlled the simulators through a generic corded computer mouse.

3.3 Experimental design

The experiment was a 2x2 completely randomized design. The independent variables were UVD type (UAV, UGV), and MCH-UVD type (general, specific). Equal number of participants was assigned to each of the four conditions. Each condition lasted 15 minutes.

3.4 Experimental tasks

Participants supervised one UAV or four UGVs. One UV and one operator is representative of many current operations, whereas the supervision of multiple UVs by one operator is the direction that UV operations are headed towards [23]. Our intent was not to compare performance across the two UV displays per se, but to compare how well the MCH-UVD scales helped identify deficiencies in different displays. Thus, experimental results were not compared across the two UV displays.

3.4.1 UAV Mission/Display

The UAV condition was a dynamic search and target acquisition mission. OPS-USERS allowed the participants to search for and track targets with a single UAV while monitoring the UAV's flight path and event timeline. The UAV was designed to search for targets and then track the targets upon target identification. Participants were instructed to monitor a water canal for passing ships. The objective was to maximize the number of ships found and the amount of time the ships were tracked.

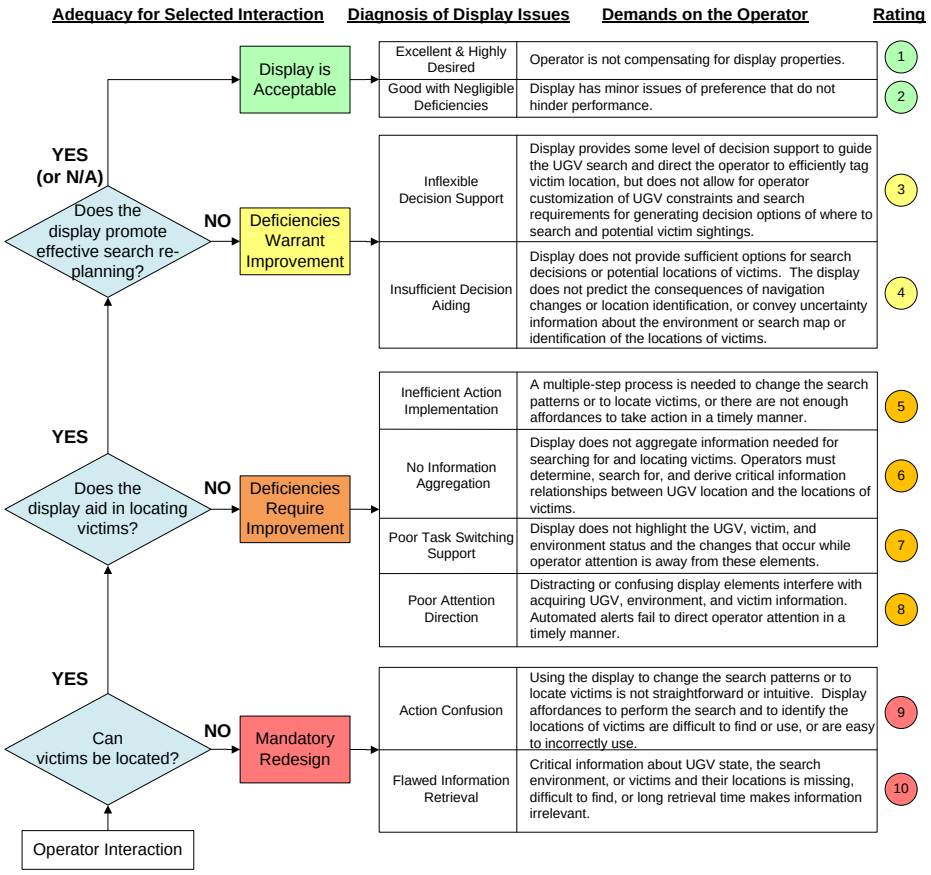

Figure 2. Specific MCH-UVD Diagnosis Tool for a UGV Search Mission.

The display (Figure 3) provided a map for tracking the UAV on its present flight path, along with an interaction control panel (bottom) for making UAV control inputs, and a panel for accepting or rejecting automation plans (left). Control inputs included creating a search or tracking task for the UAV, setting task start and end times, and creating task duration. For the purposes of this experiment, the operators were allowed to assign the UAV one task at a time. Thus, both the task value and the delay penalty seen in Figure 3 were always set to high. The operator could also modify and delete tasks using the control panel. The flight time and range of the UAV was limited by fuel, so the operator had to be aware of the fuel quantity indicated by a fuel gauge above the UAV icon. The operator had to periodically allow the UAV to return to base to refuel.

To search the canal for a target, the operator had to select a location in the canal and create a search task, and then had to accept the plan presented by the automation. Once the operator accepted the plan, a thin green line appeared showing the projected flight path from the current location of the UAV to the task, and the UAV flew to the task. Once the UAV flew over a ship (target) in the canal, the ship appeared on the map and an initial tracking task was automatically presented to the operator. The operator then had to accept the initial tracking task and the UAV tracked the ship for the duration specified by the operator.

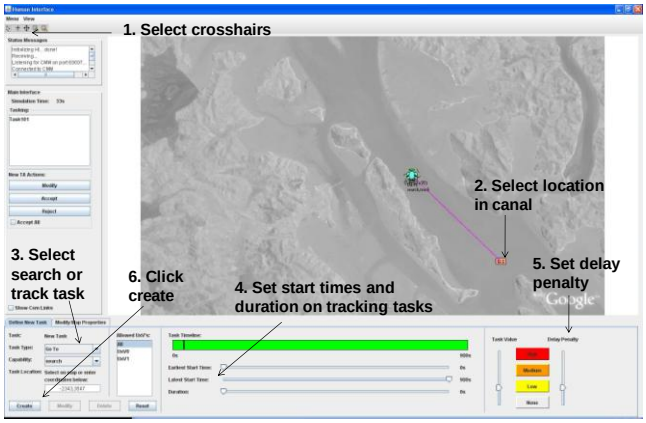

Figure 3. OPS-USERS display (Including the instructions on creating a search task).

3.4.2 UGV Mission/Display

Participants supervised four UGVs conducting a search mission for victims in a warehouse (Figure 4). The objective of the mission was to find and mark the location of as many victims as possible. The Robot List (upper left) showed each robot's camera view, whereas the Image Viewer (middle left) displayed the selected robot's camera view. Robots could be controlled by either setting way-points on the Mission panel (upper right), or by teleoperation (lower left). The Mission panel allowed operators to

create, clear, and modify waypoints. Panoramic images were taken at the terminal point of waypoint sequences, which were displayed on the Image Viewer (middle left) with operator's request. Through the teleoperation panel, it was also possible to pan and tilt the cameras. After victims were spotted in the UGV's video or in the panoramic images, operators were responsible for marking the victim's location on the Map Data Viewer (lower right frame) using a pop-up frame.

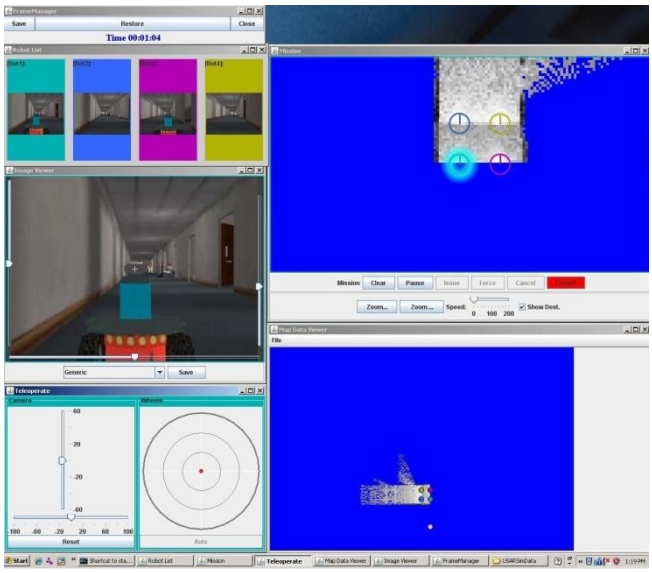

Figure 4. USARSim display.

3.4.3 Post-Mission Tasks

Following the test scenario, participants rated their workload on a Likert scale ranging from 1 to 5. Then, the test proctor showed the participants the general or specific MCH-UVD, dependent on the assigned condition, and gave a brief explanation of how to use the MCH-UVD. Participants then reported their MCH-UVD rating with an explanation for why they chose that rating and if they found any additional display deficiencies. The participants were instructed to examine all the MCH-UVD ratings before indicating these additional problems. Before compensating the participant, the test proctor asked a set of pre-determined post-test questions aimed to assess participant attitudes towards MCH-UVD.

4. RESULTS

4.1 MCH-UVD Ratings

An ordered logit model, specifically proportional odds, was developed to compare the four general groupings of ratings (i.e., display is acceptable, deficiencies warrant improvement, deficiencies require improvement, and mandatory redesign) between different displays (UAV and UGV) and MCH-UVD types (general and specific). The model was adjusted for individual differences (age and gender), subjective workload, and experimenter ratings of the operator (level of mission success, and level of display misuse). Only level of display misuse was significant and was kept in the model (α=.05).

Level of mission success and level of display misuse were assessed by the experimenter. The alternatives for level of mission success were very poor, poor, well, and very well. The experimenter followed a set of pre-determined criteria while rating mission success (e.g., focusing only on a small region of the canal while searching for ships received a "very poor" rating). The experimenter recorded any incorrect user input to the display as a misuse. The alternatives for the level of display misuse were: no, a few, some, and many misuses. Because there were only a few data points in each of the four levels, these levels were combined and collapsed into two: many or some misuses, and few or no misuses.

UVD type, MCH type, and UVD x MCH interaction were not significant. In the general MCH setting, five participants rated the UAV display to be acceptable, whereas in the specific MCH setting, there was only one participant who did the same. While not statistically significant, there is a trend that suggests that the general scale can lead to more optimistic ratings. None of the participants thought that a redesign was mandatory (ratings 9, 10).

The level of display misuse had a significant effect on MCH-UVD rating ($\chi^2(1)=9.08$, p=.003) (Table 1). Compared to having many or some misuses, having a few or no misuses increased the odds of selecting a better rating by an estimated 516% (Odds-Ratio: 6.16, 95% CI: 1.89, 20.09). This provides preliminary evidence for the validity of the scale suggesting that the order of severity indicated in the MCH-UVD is directly proportional to the level of display misuse. That is, displays that induce more operator misuses are also rated more severely in the MCH-UVD.

Table 1. Number of observations in MCH-UVD categories by level of misuse

	MCH-UVD Rating		
Level of Misuse	**Display is acceptable**	**Deficiencies warrant improvement**	**Deficiencies require improvement**
Few or None	9	25	10
Some or Many	0	6	10

The participant ratings helped us refine MCH-UVD. The scales evaluated in the experiment were earlier versions and slightly different than the scales previously discussed. Rating number seven was initially "change blindness". The experimental data showed that there were no responses in this rating level. The lack of responses for "change blindness" can be explained in part by its large overlap with rating eight, "poor attention direction", and the term "poor attention direction" being easier to understand, and thus potentially leading participants to select it rather than "change blindness". Based on this finding, rating number seven was modified from "change blindness" to "poor task switching support", which also relates to attention allocation but is more distinctive from other ratings.

4.2 Display Deficiencies Identified

This section presents the total number of UV display deficiencies that were identified by each participant across the general and specific scales. This analysis is then followed by a comparison of the total number of unique display deficiencies collectively identified by all the participants in each UVD type.

4.2.1 UAV Display

Although only marginally significant, it appeared that participants were likely to identify more UAV display deficiencies with the specific scale compared to the general scale ($\chi^2(1)=3.13$, p=.08). A large number of deficiencies identified by all participants in the general scale condition were identical to the deficiencies

identified by all participants in the specific scale condition. Four participants (27%) in each MCH condition identified a lack of display support for search guidance. A few participants indicated that better information on the remaining fuel level was necessary (N_G[1]=2, N_S=1). Participants also identified deficiencies related to path planning and re-planning. In particular, the participants indicated that time delays with respect to control inputs, automatic updating, and accepting and prioritizing tasks made it harder to use the display (N_G=5, N_S=7). One participant pointed out that the automated plan merely mirrored the route input by the operator, creating unnecessary time delays. Four participants (27%) in each MCH condition indicated that imprecise UAV paths were a problem. Participants indicated that it was difficult to change the UAV flight path (N_G=3, N_S=4).

Only one unique UAV display deficiency was identified with the general scale, which was the lack of information provided on the consequences of selected actions (N=1). There were a total of four deficiencies uniquely identified with the specific scale: the lack of information on UAV flight parameters (i.e., direction, speed), obscured duration settings, difficult target detection, and the large number of steps required to change the search patterns as well as to track targets (each by one participant).

4.2.2 UGV Display

The number of UGV display deficiencies individually identified by the participants was not significantly different between general and specific scales. Similar to the UAV display, a large number of deficiencies identified for the UGV display with the general scale were identical to the deficiencies identified with the specific scale. Out of the 15 participants in each MCH condition, the majority identified time delays to be problematic, especially the delays associated with manual control of the vehicles and the cameras, as well as the slow UV movement in general (N_G=10, N_S=7). Participants suggested having multiple cameras on a vehicle to avoid rotating the camera. One participant in each MCH type indicated that robots did not always move smoothly and follow waypoints exactly. One participant in each condition indicated that the additional step of clearing UGV paths was difficult and unnecessary. Obstacles not displayed on UGVs' paths was another deficiency identified with both the general (N_G=2) and specific scales (N_S=4). Participants thought the two maps were confusing and unnecessary (one participant in each condition). Clutter was also deemed to be a problem (N_G=1, N_S=2).

There were a total of six deficiencies uniquely identified with the general scale. These problems included the blue background (N=1), the pop-up distractions (N=1), the lack of alerts before two UGVs collided (N=1), the lack of UV idle indication (N=1), and the lack of display customizability (N=1).

There were three uniquely identified deficiencies with the specific scale. These deficiencies included the difficulty in switching between the four robots views (N=1), as well as the issues related to UGV orientation and depth perception. Specifically, the camera angle made it difficult to know how UGVs were oriented (N=3), and it was hard for participants to estimate distances on the map based on the video feed depth perception (N=6). Therefore, participants had to place several markers to get to a desired location. This display deficiency (inaccurate goal assignment) was identified by a large number (N=9) of participants with the specific scale and by no participants with the general scale, and is critical since it can significantly interfere with UV control. Thus in this case, the specific scale helped a larger number of operators identify a major display deficiency, which was not captured by the general scale.

4.3 Feedback on MCH-UVD

Out of the 57 participants who identified display deficiencies, 49, that is, 86% (N_G=23, N_S=26) thought that MCH-UVD helped them identify these display problems. 32% (NG=10, N_S=8) of the 57 participants said that they could not have recognized these deficiencies without the help of MCH-UVD. Fourteen percent (N_G=4, N_S=4) said that they could have identified deficiencies but would not be able to indicate the severity. An additional 12% (N_G=4, N_S=2) also indicated that they could have identified deficiencies but would not be able to describe them accurately.

There were mixed responses with respect to the design of the scales. The aspects of MCH-UVD categorized as being most useful included the detailed descriptions of display issues (N_G=7, N_S=14), flowchart (N_G=16, N_S=11), severity scale (N_G=1, N_S=2), and color coding (N_G=2, N_S=4). The aspects of MCH-UVD categorized as being least useful included the flowchart (N_G=9, N_S=10), technical terms (N_G=13, N_S=13), and wording being too long (N_G=3, N_S=16). Overall, the views on the usefulness of the flowchart format were split about in half. Some participants categorized the flowchart to be the most useful aspect guiding them in their ratings, whereas others thought that the flowchart questions were too broad, and led them to the wrong ratings. Nine participants (15%) suggested using checklists rather than picking one specific rating. Forty three percent (general: 21.5%, specific: 21.5%) found the language to be too technical and difficult to understand at times, and 32% (general: 6%, specific: 26%) found the wording to be too long. Twenty three percent (general: 6%, specific: 17%) suggested having more ratings for more display issues.

5. DISCUSSION

A standardized subjective display evaluation tool is an inexpensive and easy way to identify UV display improvements, as developers and testers can receive quick feedback. We proposed one such scale, MCH-UVD, in two forms: general or mission/UV specific. We also conducted a preliminary evaluation of the scale through an experiment. Participants who had more misuses with a display gave it a worse rating. Moreover, almost all of the participants (86%) thought that MCH-UVD helped them identify display deficiencies, and some (32%) said that they could not have identified the deficiencies otherwise. Although these findings are promising, it is not clear if MCH-UVD is better than other subjective methods, such as heuristic usability testing or expert evaluations. Future research should compare MCH-UVD to these existing methods.

The experiment compared the general and the UV/mission-specific scales. Although only marginally significant, individuals appeared to identify more deficiencies with the specific scale as compared to the general one. Given the limited experimental sample size, this is an important finding which has implications for the use of MCH-UVD in practice. When there are only a few operators available to rate a UV display, these results show that more deficiencies can be identified with the specific MCH-UVD.

While more unique deficiencies were identified with the general scale, these occurrences were not clustered around any clear

[1] Response number for general scale: N_G, for specific scale: N_S

problems. The unique deficiencies identified with the specific scale were clustered around a major design flaw not identified with the general scale. Thus, the likelihood that the specific scale could identify a major display deficiency is higher than with the general scale. Longitudinal data from actual practice with the scales could provide more insight on how much additional benefit the specific scale provides, and if this additional benefit is worth the effort. The amount of time required generating the specific scale and the additional benefit it may provide creates a trade-off.

About half of the participants considered the specific scale to be too wordy whereas there were only a few participants who thought the same for the general scale. An equal number of participants (43% of the total) evaluating specific and general scales indicated that some of the technical language was confusing. The level of technicality and the number of words needed to identify technical terms is a tradeoff, which has to be decided upon by the practitioners. Even if some participants considered the specific scale to be too wordy, others thought that the detailed descriptions were helpful. If operators are familiar with the technical terms, the amount of words used to explain them can be reduced.

In current operations, UVs are often designed to perform multiple missions, either singly or concurrently. An advantage of the specific scale is that it can be custom designed to consider more than one mission. Display developers can choose to administer multiple single-mission specific scales for each mission type, or combine the information of multiple missions into one specific scale to evaluate how a display supports all missions at once.

Because MCH-UVD diagnosis tools only provide one subjective measure of operator-UV interaction, other objective metrics should be collected to get a more comprehensive picture of how a UV display supports an operator in supervising a UV mission.

6. ACKNOWLEDGMENTS

This research is funded by the US Army Aberdeen Test Center (ATC). We would like to acknowledge Brooke Abounader (ATC), who provided valuable technical support and comments. Thanks to Ryne Barry and Javier Garcia for their help in data collection. Special thanks to the following persons for providing their UV displays and support: Olivier Toupet (Aurora Flight Sciences), Paul Scerri and Prasanna Velagapudi (Carnegie Mellon U.), Michael Lewis (U. of Pittsburgh), and Cameron Fraser (MIT).

7. REFERENCES

[1] DOD 2007. *Unmanned systems roadmap*. 2007-2032, Office of the Secretary of Defense, Washington, D.C.

[2] Harper, R. P. and Cooper, G. E. 1986. Handling qualities and pilot evaluation. *Journal of Guidance, Control, and Dynamics*, 9, 5, 515-529.

[3] Cummings, M. L., Myers, K. and Scott, S. D. 2006. Modified Cooper Harper evaluation tool for unmanned vehicle displays. In *Proceedings of UVS Canada: Conference on Unmanned Vehicle Systems Canada*, Montebello, PQ, Canada.

[4] Parasuraman, R., Sheridan, T. and Wickens, C. D. 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, 3, 286-297.

[5] Shneiderman, B. 1998. *Designing the User-Interface: Strategies for Effective Human-Computer Interaction*. Addison Wesley Longman, Reading, MA.

[6] Nielsen, J. 1993. *Usability Engineering*. Morgan Kaufmann, San Francisco.

[7] Brzezinski, A. S. *StarVis: A Configural Decision Support Tool for Schedule Management of Multiple Unmanned Aerial Vehicles*. Master of Science, Massachusetts Institute of Technology, Cambridge, MA, 2008.

[8] Rubinstein, T. and Mason, A. F. 1979. The accident that shouldn't have happened: an analysis of the Three Mile Island. *IEEE Spectrum*, November, 33-57.

[9] Norman, D. A. 2002. *The Design of Everyday Things*. Basic Books, New York.

[10] Billings, C. E. 1997. *Aviation Automation: The Search for a Human Centered Approach*. Lawrence Erlbaum Associates, Mahwah, NJ.

[11] Carrigan, G. P., Long, D., Cummings, M. L. and Duffner, J. 2008. Human factors analysis of predator B crash. In *Proceedings of AUVSI: Unmanned Systems North America*, San Diego, CA.

[12] NTSB 1995. *Marine Accident Report, Grounding of the Panamian Passenger Ship Royal Majesty on Rose and Crown Shoal near Nantucket, Massachusetts*. NTSB/MAR-97/01, National Transportation Safety Board.

[13] Sheridan, T. B. 1972. On how often the operator supervisor should sample. *IEEE Transactions on Systems, Science, and Cybernetics*, SSC-6, 140-145.

[14] Rogers, R. D. and Monsell, S. 1995. Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124, 2, 207-231.

[15] McFarlane, D. C. 2002. Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction*, 17, 1, 63-139.

[16] Degani, A. and Wiener, E. L. 1990. *Human factors of flight-deck checklists: the normal checklist*. NASA Contractor Report 177549, NASA Ames Research Center.

[17] Wickens, C. D. and Carswell, C. M. 1995. The proximity compatibility principle: Its psychological foundation and its relevance to display design. *Human Factors*, 37, 3, 473-494.

[18] Woods, D. D. 1995. Toward a theoretical base for representation design in the computer medium: Ecological perception and aiding human cognition. In J. M. Flach, P. A. Hancock, J. Caird and K. J. Vicente (Eds.), *An Ecological Approach to Human-Machine Systems I: A Global Perspective*. Lawrence Erlbaum Associates, Hillsdale, N.J., 157-188.

[19] Cummings, M. L. and Mitchell, P. J. 2006. Automated scheduling decision support for supervisory control of multiple UAVs. *AIAA Journal of Aerospace Computing, Information, and Communication*, 3, 6, 294-308.

[20] Cummings, M. L., Brzezinski, A. S. and Lee, J. D. 2007. The impact of intelligent aiding for multiple unmanned aerial vehicle schedule management. *IEEE Intelligent Systems: Special Issue on Interacting with Autonomy*, 22, 2, 52-59.

[21] Lewis, M., Wang, J. and Hughes, S. 2007. USARsim: Simulation for the study of human-robot interaction. *Journal of Cognitive Engineering and Decision Making*, 1, 1, 98-120.

[22] Wang, J. and Lewis, M. 2007. Human control of cooperating robot teams. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, Arlington, VA.

[23] Cummings, M. L., Bruni, S., Mercier, S. and Mitchell, P. J. 2007. Automation architecture for single operator, multiple UAV command and control. *The International Command and Control Journal*, 1, 2, 1-24.