Investigation of the sensitivity of physiological, performance, and subjective measures for identifying changes in novice intelligence analyst workload

Lisa Tripp, Robert Nelson, Elliot Humphrey, Chad Tossell
Air Force Research Laboratory
WPAFB, OH
Lisa.tripp.1@us.af.mil,
Robert.Nelson44@us.af.mil,
Elliot.Humphrey.1@us.af.mil,
Chad.Tossell@us.af.mil

Jennifer Winner, Jerred Holt
Lumir Research Institute
WPAFB, OH
Jennifer.Winner.ctr@us.af.mil,
Jerred.Holt.ctr@us.af.mil

ABSTRACT

The United States Air Force has a vested interest in advancing intelligence, surveillance, and reconnaissance technologies. Although software and hardware testing is performed for these technologies to demonstrate functionality, only limited research has investigated the effect of these tools on human performance. This research describes a process for the identification of suitable metrics to assess the effectiveness of new ISR technologies. We used several factors to determine the potential suitability of candidate measures including their relative sensitivity, reliability, content validity, and task intrusiveness. Additionally, the sensitivity of several measures, including performance-based, physiological and subjective measures, for the discrimination between levels of difficulty of imagery analyst tasking were compared. Twenty participants from a school for training intelligence analysts volunteered. Real recorded footage from two imagery types, wide area motion imagery and full motion video, was presented to analysts in short video clips. Tasking for each clip was provided prior to viewing. Tasking was developed by a subject matter expert and validated by five career analysts who independently rated the tasking in terms of difficulty. Performance data showed a significant difference based on difficulty of tasking as predicted ($F(1,19) = 220.32$, $p < .001$), as did subjective difficulty ratings assessed by the NASA-Task Load Index ($F(1,19) = 12.84$, $p < .01$). The sensitivity of physiological data to difficulty was mixed. Significant differences based on difficulty rating were identified for fixation duration ($F(1, 14) = 5.30$, $p = .037$) and saccade duration ($F(1, 14) = 15.13$, $p < .01$). However, no significant differences were identified in heart rate or heart rate variability ($p > .05$). There were also no significant differences in indices of workload across imagery types. The suitability and applications of these measures for assessing intelligence analyst performance in simulated analyst operational environments is discussed.

ABOUT THE AUTHORS

Lisa Tripp is the ISR Training Research lead for the Air Force Research Laboratory. She received her Ph.D. in Experimental Psychology from Washington State University.

Robert Nelson is the program manager for command and control training research at Air Force Research Laboratory. He has a B.S in Psychology from Utah State University.

Elliot Humphrey is the ISR Training Research Project Lead at the Air Force Research Laboratory. He has a B.S. in Psychology from Bethel University.

Chad Tossell directs the C4ISR Training Research for the Warfighter Readiness Research Division. He has a Ph.D. in Psychology from Rice University.

Jennifer Winner is a Lead Research Scientist with Lumir. She has a M.S. in Psychology from Arizona State University.

Jerrod Holt is a Research Scientist with Lumir. He has a M.S. in Psychology from Wright State University.

Investigation of the sensitivity of physiological, performance, and subjective measures for identifying changes in novice intelligence analyst workload

Lisa Tripp, Robert Nelson, Elliot Humphrey, Chad Tossell
Air Force Research Laboratory
WPAFB, OH
Lisa.tripp.1@us.af.mil,
Robert.Nelson44@us.af.mil,
Elliot.Humphrey.1@us.af.mil,
Chad.Tossell@us.af.mil

Jennifer Winner, Jerred Holt
Lumir Research Institute
WPAFB, OH
Jennifer.Winner.ctr@us.af.mil,
Jerred.Holt.ctr@us.af.mil

INTRODUCTION

The United States Air Force (USAF) has a vested interest in advancing intelligence, surveillance, and reconnaissance (ISR) technologies, and the capability to rapidly and effectively leverage the information obtained by these technologies. ISR capabilities play a critical role in the USAF's warfighting capabilities, and over the past decade the USAF has procured a variety of new sensors and platforms that have enabled significant gains in the availability of imagery to inform decision makers during combat operations. The changes in ISR capabilities operated by the USAF have required substantial changes in the technology used to process, analyze, and disseminate information. Capabilities were rapidly developed and readily accepted by the community that was desperate to keep up with emerging sensor capabilities and evolving requirements. While the development of these capabilities undoubtedly aided the Air Force in meeting their mission objectives, unfortunately, many of the capabilities provided only short-term solutions, had high levels of redundancy with other capabilities, contained unpredicted bugs, and had limited interoperability with standard Air Force systems.

What capabilities does the USAF have for identifying tools that are effective? In other words, what processes and capabilities does the USAF have to be an informed consumer with regard to technologies for intelligence analysts? In the current fiscally-constrained environment, it is paramount that selection of new capabilities be informed by data. As part of this initiative, a research environment was developed to evaluate new tools for ISR analysis in a human-in-the-loop, simulated operational environment, providing objective and subjective empirical data to inform decision makers. This led to the development of the Analyst Test Bed (ATB), a joint collaboration between the Air Force Research Laboratory and Alliance for the Human Effectiveness and Advancement (AHEAD). This paper will describe the process leveraged for the identification of suitable metrics to assess the effectiveness of new ISR technologies in a realistic, simulated operational environment. In it we present a study aimed at equipping the ATB with measurement capabilities to quantitatively and qualitatively assess the impact and effectiveness of new analysis tools on analyst performance. We integrated a variety of behavioral and physiological measures and evaluated them within the context of an intelligence analysis environment. This was the first step in developing a capability for the empirical assessment of human-centered operator performance for intelligence analysis.

Development, selection, and integration of metrics for assessing human analyst performance in simulated operational environments are significant challenges, especially when attempting to be an ecologically valid test environment. Controlled laboratory research can provide the capability to make causal inferences between intervention and outcome and yield high internal validity (e.g., Anderson & Bushman, 1997). On the other hand, these causal inferences may not exhibit the same relationships when other factors are present (i.e., limited generalizability or low external validity), as they are in more naturalistic environments. Naturalistic observation allows for studying a phenomenon in a setting where all potential factors are present, but this lack of control inhibits causal inference. Although, it should be noted that some research has found similar results in observational studies as very controlled studies. For instance, in the medical domain, research comparing randomized, controlled trials and observational studies found negligible difference between confidence intervals generated via the two methods (e.g., Concato, Shah, & Horwitz, 2006). One goal for this effort was to maximize generalizability to the operational

environment to the extent possible while still gathering empirical data (i.e., metrics) to assess the effectiveness of the tools in realistic environments.

Another purpose of this research was to verify that the capability exists to experimentally impose realistic tasking with sufficient fidelity to result in increases in workload, and additionally, to establish that the capability exists to detect these differences via the current performance-based, physiological, and subjective measurement capabilities. Performance metrics were identified through a multifaceted approach to user analysis (i.e., Mission Essential Competency; Bennett et al., 2007). This process allowed us to identify tasking perceived by experienced analysts to impose increased workload.

The NASA-Task Load Index (TLX) is a multi-dimensional scale used to measure workload. It is validated, sensitive to changes in workload, and has high diagnosticity (Rubio et al., 2004; Hill et al., 1992). Furthermore, it has been used in a variety of applied and academic settings; and is one of the most often-used measures for identifying changes in workload (Hill, 1992; Hart, 2006). It has been shown to have greater sensitivity, concurrent validity with performance, and diagnosticity when compared to other measures of workload (Rubio et al., 2004).

Ocular behavioral data has also been shown to correlate with task difficulty and cognitive workload. (Dahlstrom et al., 2011; Palinko & Kun, 2011; Pomplun & Sunkara, 2003). As task complexity increases in a simulated air-traffic control task, blink duration and saccade distance were found to significantly decrease. In addition, pupil dilation was significantly greater with increases in workload (Ahlstrom & Friedman-Berg, 2006). Increased pupil dilation is a reliable and involuntary response associated with short and long term memory access, mental arithmetic, reading comprehension, vigilance, and perceptual tasks (Klingner et al., 2008). In a simulated driving task, Palinko et al. (2010) found that mean pupil diameter changed significantly and positively correlated with increased cognitive workload.

Heart-based physiological metrics have been shown to correlate with task difficulty and workload. Corresponding changes in heart rate were documented (increases for higher workload flight segments and decreases for lower workload flight segments of simulated and actual flights) for aviation trainees and pilots (Dahlstrom & Nahlinder, 2009; Dahlstrom et al., 2011). Parsons et al. (2009) found that participants' median heartbeats per minute (BPM) were significantly higher when directly interacting with virtual environments rather than passively observing the same virtual environments, indicating that higher levels of immersion correlate with increased physiological reactions. We predict that task difficulty will be rated high by SMEs for video snippets with higher perceptual load, greater similarity between targets and distractors, and a larger number of occlusions. Furthermore, it is predicted that these video snippets will increase performance-based (accuracy and time), physiological, and subjective measures of workload.

METHODOLOGY

Participants

Twenty participants (6F, 14M**)** were recruited from the Advanced Technical Intelligence Center (ATIC) in Dayton, OH. All participants were current students or alumni of ATIC. The average age for the 19 participants reporting demographic data was 42 years (age range: 20 - 66). One person declined to report an age. Twelve of the participants had taken or were currently enrolled in the ATIC Analyst Bootcamp course, six had completed basic military training (BMT), five reported having had geospatial intelligence training, and two reported a class on full motion video. Participants reported on other relevant training courses, including ATIC Advanced Technical Intelligence (3), Security Forces Training School (1), SOCET GXP Seminars (1), U.S. Army Military Police (1), and U.S. Army Calvary Scout (1). Two participants reported that their experiences included the National Air & Space Intelligence Center (NASIC), while no other relevant experiences were reported (e.g., AF DCGS-A, Army Intel Brigade, NGA Imagery, MQ-1/9, Real-time FMV in DGS/NASIC/NGA). Four participants reported previous deployments.

Stimuli

The stimuli were derived from two types of imagery: real-world full motion imagery (FMV) and real-world Wide Area Motion Imagery (WAMI). The FMV leveraged was high definition, color video footage. The WAMI was

lower-resolution black and white footage captured at 60Hz. Since this is real footage and hence varies on a variety of dimensions, sources of difficulty for these tasks were identified through cognitive task analysis with subject matter experts (SMEs). A subject matter expert (SME), an intelligence analyst with over 20 years of experience in the field, was asked to identify appropriate tasking for 40 video snippets such that 20 videos would be relatively easy for a novice analyst and 20 would be difficult. After the creation of the stimuli by a single SME, the video snippets were then independently rated by five additional SMEs for difficulty and using these ratings were used to categorize the scenarios dichotomously as either easy or difficult. The top 25% and bottom 25% were determined to have high and low levels of difficulty and were selected for this experiment. Difficulty ratings were collected using a 5-point Likert scale with an average rating for difficult scenarios of 1.8 and an average rating for easy scenarios of 3.4. The average variability among raters for the difficulty of selected snippets was .8 on a 5-point Likert scale. The duration of each video clip was 60 seconds.

Tasks

Participants were asked to perform tasking known to be common in the field of imagery analysts (e.g., slant count) using prerecorded FMV or WAMI. Participants were asked to individually respond to the tasking following each trial using pencil and paper.

Experimental Design

The experiment leveraged a 2x2 factorial design with factors: task difficulty (easy vs. difficult) and imagery type (FMV vs. WAMI).

Table 1: Independent Variables (IVs) and Dependent Variables (DVs)

IVs	DVs
Task difficulty (easy; difficult)	Performance (mean counting deviation)
Imagery Type (FMV;WAMI)	Subjective (rating; NASA-TLX)
	Physiological (eye-tracking; ECG)

The video clips were grouped into four 5 minute blocks (i.e., two 'easy' blocks and two 'difficult' blocks) with a five minute break between Block 2 and Block 3. The order of the blocks was counterbalanced. The measures collected during each block of trials include task performance (i.e., accuracy of response to tasking), subjective workload (i.e., NASA-TLX), and physiological measures of workload (i.e., heart and eye tracking data). Perceived task difficulty was collected from participants using the same 5 point Likert scale administered to SMEs prior to data collection. The difficulty scale was filled out after each video was viewed. Task performance was measured by calculating the difference value between the participants count and the correct count.

Oculomotor and pupillary dynamics were monitored and recorded using a Smart Eye Pro 5.6 (120Hz) eye-tracking system (Smart Eye AB – Göteborg, Sweden). In real time, Smart Eye extracted 23 parameters including the following: gaze direction, gaze original, pupil diameter, eye lid opening, blink rate, and fixation duration. The experimenters sought to achieve +/- 2-degrees of accuracy for each calibration point.

Heart rate information was captured using the Equivital™ sensor. Equivital™ captures full electrocardiographic (ECG) data as well as heart rate and breath rate. For the purposes of this study, we will be looking at interbeat intervals, heart rate, and heart rate variability.

Procedures

Upon arrival, participants received informed consent documents and were fitted for the Equivital™ sensor. After calibration of the eye tracking system, participants were provided imagery examples (still pictures) to familiarize them with the imagery they would be viewing throughout the trials. Imagery content was labeled to provide examples of what various objects might look like in the two imagery types (e.g., pedestrians, bicyclists, motorized bikes, automobiles). Participants were then provided tasking which mimicked what they would see in the experimental trials. Specifically, they were provided guidance as to the area within the imagery that they were to monitor and provided instructions as to the type of activity that they were to report on. For example, 'Count the

number of people entering the building indicated'. After completing the training trials, participants were asked to sit quietly for five minutes without any tasking. The purpose of this break was to allow a five-minute resting period for observation of the physiological data (heart rate) to establish a baseline level for use in the data analysis.

Participants completed four blocks consisting of five trials each. Videos for each trial were 60 seconds in length. The imagery and associated tasking were implemented in a counter-balanced distribution of the imagery (i.e. FMV or WAMI) and task difficulty (i.e., easy or difficult), which was determined in advance. Participants were assigned a block order upon arrival at the laboratory.

For each trial, participants received advance instructions as to the area of responsibility (AOR) for the particular trial and precise instructions for the counting task (e.g., count all the motorized bikes in the area). Following the end of the video, participants recorded their count based on the imagery and provided a subjective rating.

Each block of five trials was followed immediately by administration of the NASA Task Load Index (TLX). After the second NASA TLX administration (after block 2), participants were given a second five-minute resting period for observation of the physiological data (heart rate). After the final block and administration of the NASA TLX, participants were debriefed and escorted to the changing room to remove the Equivital^{TM} vest and sensor.

Results

One individual's data was lost due to technical issues. The data for each remaining participant was averaged across all the available trials of a given condition. In the following sections we present the results for counting performance as well as those for subjective workload ratings, and eye tracking measures. No significant differences in heart rate nor heart rate variability were identified (p >.05) across difficulty or task conditions; no further analyses on these measures will be presented. All analyses were performed with sphericity assumed unless otherwise noted.

Counting Performance

Analysis of the counting performance data show a significant interaction between expert difficulty categorization and imagery type, $F(1,19) = 12.60$, $p = .002$, partial eta squared = .399 (Figure 1). Participants were more accurate in their counts for easy tasking paired with FMV than WAMI; however, for difficult tasking, participants were more accurate with WAMI than for FMV.

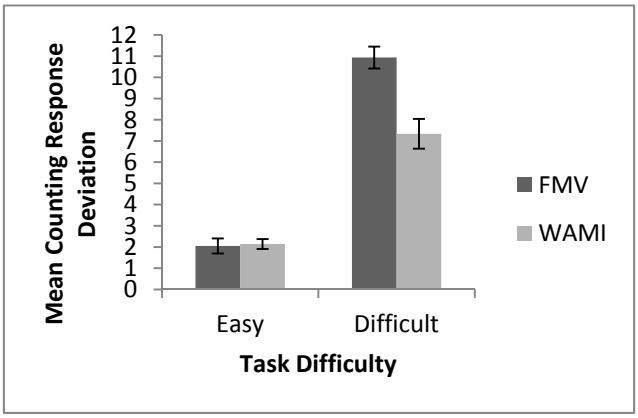

Figure 1: Mean Counting Response Deviation by Imagery Type and Task Difficulty

Subjective Ratings of Task Difficulty

There was a significant difference in novice difficulty ratings based on the expert categorization of task difficulty $F(1,19) = 66.48$, $p < .001$, partial eta squared .778 (Figure 2). There was no evidence that imagery type resulted in differences in difficulty ratings $p > .05$. There was no evidence of an interaction between expert difficulty categorization and imagery type, $p > .05$.

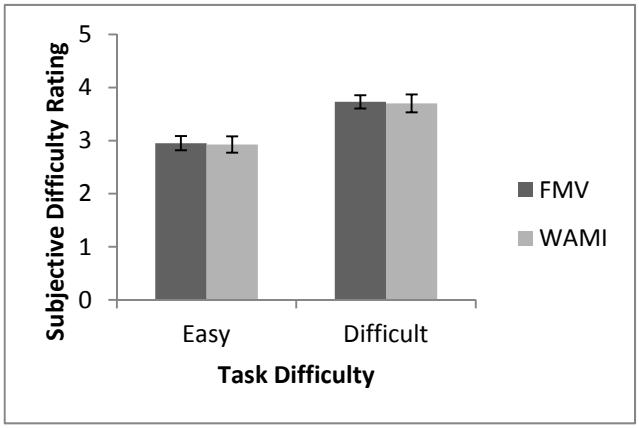

Figure 2: Mean Subjective Difficulty Rating by Imagery Type and Task Difficulty

Subjective Workload Ratings

Analyses were conducted for the overall workload assessment. There was a significant difference in novices' self-reported workload based on the expert categorization of difficulty $F(1,19) = 12.84$, $p < .01$, partial eta squared .403 (Figure 3). There was no evidence that imagery type resulted in differences in self-reported workload, $p > 0.5$. There was no

evidence of an interaction between expert difficulty categorization and imagery type, $p > .05$. Analyses were also conducted for each of the subscales. On the frustration, performance, physical and temporal subscales was no significant difference in novices' self-reported workload based on the expert categorization of difficulty, no significant differences in self-reported workload, and no significant interaction, $p > 0.5$. Significant results are detailed in the following paragraphs.

There was a significant difference in novices' self-reported effort subscale based on the expert categorization of difficulty $F(1,19) = 7.53$, $p < .02$, partial eta squared .284 (Table 2). There was no evidence that imagery type resulted in differences in self-reported effort, $p > .05$. There was no evidence of an interaction between expert difficulty categorization and imagery type, $p > .5$.

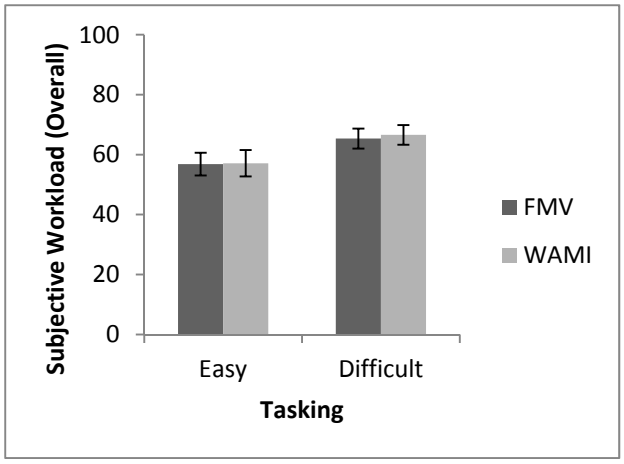

Figure 3: Overall Subjective Workload by Imagery Type and Task Difficulty

Table 2. Mean Subjective Effort Ratings by Condition

	Mean	St. Dev	N	SE
Easy, FMV	9.7667	6.39819	20	1.430679
Easy, WAMI	10.3667	7.16628	20	1.602429
Difficult, FMV	13.35	7.34487	20	1.642363
Difficult, WAMI	14.9667	8.82043	20	1.972308

There was a significant difference in novices' self-reported mental demand subscale based on the expert categorization of difficulty $F(1,19) = 4.81$, $p = .041$, partial eta squared .202 (Table 3). There was no evidence that imagery type resulted in differences in self-reported mental demand, $p > .05$. There was no evidence of an interaction between expert difficulty categorization and imagery type, $p > .05$.

Table 3. Mean Subjective Mental Demand Ratings by Condition

	Mean	St. Dev	N	SE
Easy, FMV	17.7167	7.84296	20	1.753739
Easy, WAMI	17.5833	9.06111	20	2.026126
Difficult, FMV	21.6833	7.66207	20	1.713291
Difficult, WAMI	18.9833	6.50728	20	1.455072

Eye tracking

The data for each eye-tracking measure listed in Table 4 were analyzed using a 2 (sensor – WAMI vs. FMV) X 2 (load – low vs. high) repeated measures ANOVA. The Greenhouse-Geiser correction for sphericity was considered but ultimately had no effect on the results given that each condition had only two levels. Because the goals of the current study were to validate task manipulations and measures that have been reported extensively in the human factors literature and to select the most sensitive and ideal physiological measures from a range of candidate measures, we chose to adopt a relatively liberal statistical threshold (a standard alpha threshold of $p < 0.05$, uncorrected for multiple comparisons) due to the large number of statistical comparisons required to achieve these goals.

Results of all 2 x 2 repeated measures ANOVAs on the eye-tracking data are reported in Table 4, with plots of the condition means for all measures showing significant (Figure 4) main effects of load.

Table 4. ANOVA results of the eye-tracking data statistical analyses

	Load			Sensor			Load x Sensor		
DV	F	*p*	MSE	F	*p*	MSE	F	*p*	MSE
Fixation Duration	5.30	0.037	393.7	17.69	0.001	321.3	2.42	0.142	444.9
Saccade Duration	15.13	0.002	33.8	0.00	0.999	16.0	0.13	0.910	18.4
Blink Duration	4.62	0.050	5745.0	0.82	0.380	317.2	0.35	0.565	202.1
Pupil Diameter	4.68	0.048	2.26×10^{-7}	2.68	0.124	1.57×10^{-7}	0.01	0.911	1.81×10^{-7}

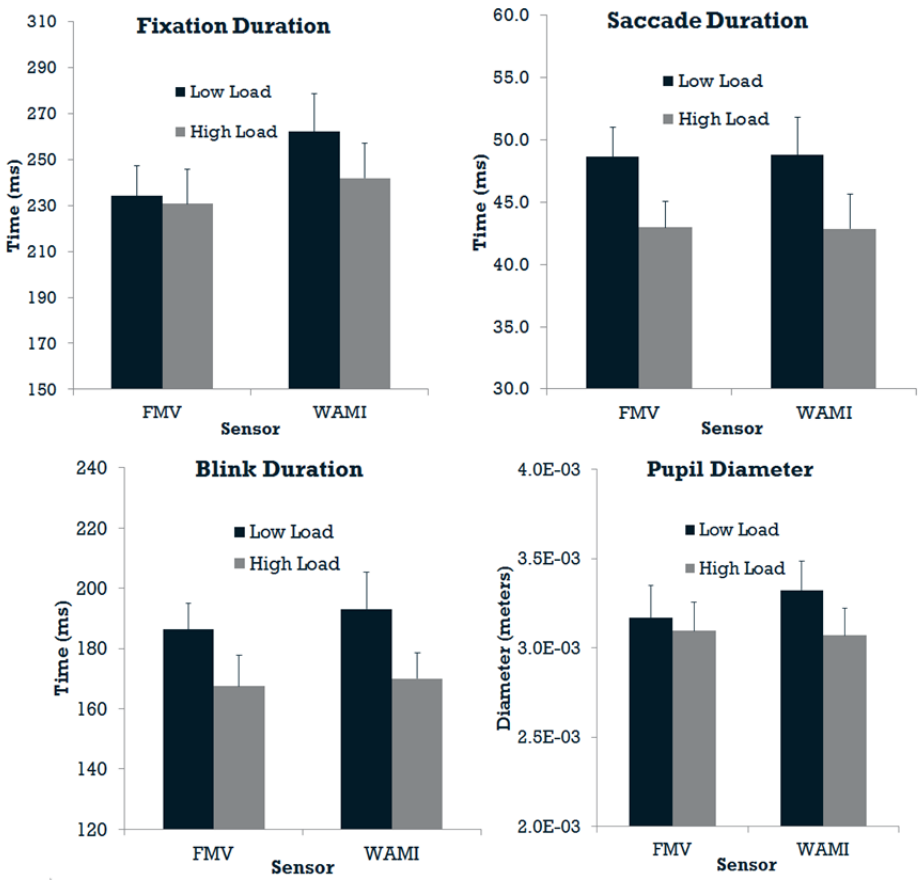

Figure 4. Plots of condition means for the measures showing a significant main effect of load at uncorrected $p < 0.05$.

DISCUSSION

Two of the primary objectives of this study were: (1) to validate that the task difficulty manipulations used here can reliably increase operator workload and that the behavioral and subjective rating data is sensitive enough to detect changes in even this type of applied environment, and (2) to validate that the non-invasive physiological sensor systems used in this study are sensitive to changes in operator workload. Regarding the first objective, the behavioral performance and subjective rating data support the hypothesis that the task manipulations significantly altered operator workload. Significant differences in novice accuracy and subjective difficulty rating corresponded with expert ratings of difficulty. Interestingly, there was no significant difference in either accuracy or subjective difficulty between tasking using the two imagery types.

The finding is surprising because the subject matter experts indicated that, in general, WAMI sensor data is significantly more difficult to work with due to its lower resolution and temporal sampling rates. There are several potential explanations. It might be that in general working with WAMI imagery does not increase analyst workload compared to FMV. However, previous data gathered using Cognitive Task Analysis methods indicates analysts at least subjectively feel that working with WAMI increases workload. It would be expected that we should at least have identified differences via the NASA-TLX. It is possible that there was not sufficient power to detect an effect between the imagery types (i.e., the effect was small). However, the data does not even indicate a trend in the anticipated direction.

The final hypothesis is that the expert analyst that identified the tasking unknowingly created tasking of different difficulty depending on the imagery type (i.e., easier tasking for WAMI than for FMV) negating the influence of imagery difficulty. Despite efforts to keep tasking difficulty equivalent across conditions, the real world imagery available may have been more conducive to more difficult tasking for FMV as compared with WAMI. Future research to compare imagery types should identify the same footage for each sensor type and use identical tasking. While this was considered prior to running the experiment, this was not possible in our design due to the availability of imagery at the time this research was conducted. The results here regarding imagery type should be interpreted cautiously.

Regarding the second objective, the data from the SmartEye ™ eye-tracker and Equivital™ physiological sensor systems will be discussed here individually, beginning with the SmartEye ™ data, before addressing the battery of physiological assessment measures as a whole. Overall, a number of measures extracted from the eye-tracking data differed significantly across the task load conditions, indicating that the eye-tracking data was sensitive to differences in workload demands. At the same time, the eye-tracking measures were largely unaffected by the type of sensor (WAMI vs. FMV) used by the participants (fixation duration and pupil quality were the only two measures that were either marginally significant or significant). This latter finding is somewhat surprising due to the fact that the study did not attempt to control for physical stimulation parameters of the two types of sensor feeds (such as luminance and contrast) as might be done in a tightly controlled basic research experiment and that luminance is known to affect eye-movement parameters such as pupil diameter. On the other hand, the pupil also expands and contracts dynamically over a matter of seconds in response to perceptual and cognitive events (Beatty, 1982), and it is possible that the pupillary dynamics evoked by the task masked any effects that physical stimulation parameters might have had on the eye-tracking measures.

While a number of measures extracted from the eye-tracking data were *sensitive* to differences in workload (i.e. they differed significantly between the two task difficulty levels), the degree to which each measure was *diagnostic* (i.e. how consistent the *direction* of each effect was with previously reported findings in the literature) was somewhat mixed. The blink-related measures extracted from the data were quite consistent with the existing body of literature. As has been reported previously, blinks were shorter in duration (Fournier, Wilson, & Swain, 1999; Veltman & Gaillard, 1998; Zheng et al, 2013) and were marginally less frequent (Fournier, Wilson, & Swain, 1999; Recarte et al, 2008; van Orden et al, 2001) under the higher workload conditions. Furthermore, the inter-blink interval was marginally smaller under the higher workload conditions, consistent with studies showing a similar effect as a function of memory load (Veltman & Gaillard, 1998). The fixation duration effect was also consistent with a study showing shorter fixation durations among anesthetists when managing a critical incident compared to scenarios lacking a critical incident (Schulz et al, 2011). However, the expected effect of increased workload on fixation duration appears to depend on what is driving workload – Zelinsky and Sheinberg (1997) found *longer* fixation durations in a basic research task when participants had to search through a screen with a larger number of elements.

Other measures exhibiting a significant effect of workload, however, indicated an effect that was not consistent with what might be predicted based on findings from the literature. The most prominent example is the significant *decrease* in pupil diameter observed here under the high workload conditions. Contrary to this finding, the majority of studies report that pupil diameter *increases* under higher workload conditions (e.g. Beatty, 1982). However, there is some evidence to suggest that the relationship between workload and pupil diameter may not be monotonic, which might explain why the effect of load on pupil diameter was reversed in the current study. Van Gerven et al (2004) tracked pupil diameter as a function of memory load across a number of load values and found that although pupil diameter increased for most step-wise increases in memory load, at the very highest level of memory load, pupil diameter decreased substantially. Similarly, another study indicated that adding a secondary task onto a primary task increased pupil diameter when the primary task was lower in mental workload, but it decreased pupil

diameter when the primary task was already higher in mental workload (Recarte et al, 2008). However, one caveat to the workload effect on pupil diameter reported here is that the effects of workload on SmartEye's pupil quality and iris loss measures –measures of signal quality – were marginally significant, with both indicating poorer signal quality under the high load conditions than under the low load conditions (see Table 4 and Figure 4). Furthermore, the effect of sensor type on pupil quality was also marginally significant, with signal quality better for the WAMI than for the FMV sensor conditions. The effect of load on pupil quality could be explained by participants moving their heads either more often or closer to the screen and out of optimal sensor calibration range in the high load condition, although this explanation is speculative. The pupil diameter measure included in the analysis here was drawn from SmartEye's Filtered Pupil Diameter parameter, which adds temporal smoothing to the current sample's diameter estimate in inverse proportion to the current pupil quality estimate. While the filtering algorithm should have reduced the impact of the effect of signal quality on pupil diameter, the main effect of load on the pupil quality indicates that simple measurement differences cannot be ruled out as the main driver of the observed workload effects on pupil diameter.

Other measures that were significantly or marginally significantly affected by workload in the current analyses have not shown this effect consistently in previous studies. In particular, recent workload studies reporting saccade duration as a dependent measure failed to find a significant or marginally significant effect of workload (Di Stasi et al, 2010; Halverson et al, 2012). A different, but related, measure – saccade distance – also has not consistently been affected by workload across studies (Halverson et al, 2012 and Schulz et al 2011 for negative results; van Orden et al, 2001 and Zelinsky & Sheinberg, 1997 for positive results). Notably, the two studies that did report effects on saccade distance (van Orden et al, 2001 and Zelinsky & Sheinberg, 1997) both manipulated overall workload by increasing the perceptual processing demands of the task. This manipulation was similar in nature to the workload manipulation used in the current effort, and the direction of the effect reported here (shorter duration saccades) is consistent with what one would expect to see in saccades of a shorter distance, as reported in van Orden et al (2001) and Zelinsky and Sheinberg (1997). Therefore, the effect of workload on saccade duration here is consistent with some reports in the literature, but it may only hold for high workload conditions associated with the perceptual difficulty of a task. In contrast, the current effort failed to find an effect of workload on eyelid opening, which contrasts somewhat with a recent claim that PERCLOSE (percentage eyelid closure, a related measure) discriminates levels of workload well (Halverson et al, 2012). Our finding is more consistent with the larger body of literature on workload. The finding reported by Halverson et al (2012) may reflect combined effects of workload and vigilance – operators performed the task for 40 minutes straight, alternating between low- and high-levels of workload in 5-minute segments. Workload and vigilance are not entirely independent constructs (Caggiano & Parasuraman, 2004; Parasuraman, 1979; Warm, Parasuraman, & Matthews, 2008), and the effect of workload on PERCLOSE may have emerged from a workload x vigilance interaction during later segments of the 40-minute sessions.

Noise in both the heart rate and eye tracking data was a substantial issue that must be considered in future extensions of this work, particularly the instances in which the signal quality metrics were affected by the workload manipulation. In the current study, several steps were taken to address these potential noise issues: (1) trials for each condition were averaged together for all participants, and the data were analyzed using a random effects model on the group-wise data, (2) trial-wise data were normalized to a per-60-second time window based on the segments of the trial in which the relevant eye-tracker signal was adequate, and (3) thresholds for individual measure instances (e.g. individual fixations and saccades) were based on the measure duration to ensure that the automated measure extraction algorithms were identifying true measure instances.

Overall, this study provided evidence for both the sensitivity and the diagnosticity of a subset of the measures extracted from a variety of data sources in relation to the workload demands of a task. As expected, behavioral performance and subjective data showed robust differences even in simulated real-world environments. For future studies, blink-related measures seem to be the most robust and diagnostic measure captured by the SmartEye ™ data, followed by saccade duration and fixation duration. For the type of tasking imagery analysts generally performed, heart rate and heart rate variability were not sensitive enough to detect differences. This may have been a function of the system selected or the environment. Note that the system selected was identified on the basis of being relatively non-disruptive to typical tasking. Additional foundational research should be conducted before leveraging electrocardiogram-based measures for diagnosticity of workload for imagery analysis. Overall, this set of converging metrics will allow us to identify changes in analyst workload in high fidelity simulated environments where missing data is common. This affords the capability to allow analysts to participate in team interactions, move

around in the environment, and act naturally while providing real-time seamless and non-disruptive data collection. While we would not recommend these conditions for answering basic research questions, these types of applied environments are critical for providing empirical data so that the USAF is an informed consumer with regard to technologies for intelligence analysts.

ACKNOWLEDGEMENTS

We'd like to recognize Dr. Daniel M Caggiano (previously with Aptima) for his support in analyzing and interpreting the eye-tracking data. We'd also like to thank the ATIC staff for the phenomenal support from the personnel and facility. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

REFERENCES

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*(7), 623-636.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91, 276-292.

Bennett W. (2007). *Understanding Mission Essential Competencies as a Work Analysis Method.* Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a474546.pdf

Caggiano, D. M., & Parasuraman, R. (2004). The role of memory representation in the vigilance decrement. *Psychonomic bulletin & review*, *11*(5), 932-937.

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, *342*(25), 1887-1892.

Dahlstrom , L.I. (2011). Effect of Nesiritide in Patients with Acute Decompensated Heart Failure. *New England Journal of Medicine*, 365(16), 1547-1547.

Dahlstrom, N., & Nahlinder, S. (2009). Mental workload in aircraft and simulator during basic civil aviation training. *The International journal of aviation psychology*, *19*(4), 309-325.

Di Stasi, L. L., Marchitto, M., Antolí, A., Baccino, T., & Cañas, J. J. (2010). Approximation of on-line mental workload index in ATC simulated multitasks. *Journal of Air Transport Management*, *16*(6), 330-333.

Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training. *International Journal of Psychophysiology*, *31*(2), 129-145.

Halverson T. (2012). Classifying Workload with Eye Movement in a Complex Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56(1), 168-172.*

Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 9, pp. 904-908). Sage Publications

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklade, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(4), 429-439.

Klingner, J., Kumar, R., & Hanrahan, P. (2008, March). Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 69-72). ACM.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010, March). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 141-144). ACM.

Palinko, O., & Kun, A. L. (2011). Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. *Proceedings of Driving Assessment*.

Parasuraman, R. (1979). Memory load and event rate control sensitivity decrements in sustained attention. *Science, 205*, 924–927.

Parsons, T. D., Courtney, C., Cosand, L., Iyer, A., Rizzo, A. A., & Oie, K. (2009). Assessment of psychophysiological differences of west point cadets and civilian controls immersed within a virtual environment. In *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience* (pp. 514-523). Springer Berlin Heidelberg.

Pomplun, M., & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of the International Conference on HCI*.

Pérez, E., Nunes, L. M., Conchillo, Á., & Recarte, M. Á. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *The Spanish Journal of Psychology, 11*(2), 374-385.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology*, *53*(1), 61-86.

Schulz, M.J. (2011). Rationale and study design of PROVHILO – *a worldwide multicenter randomized controlled trial on protective ventilation during general anesthesia for open abdominal surgery.* Received from http://www.biomedcentral.com/content/pdf/1745-6215-12-111.pdf

Van Gerven, P.M. (2004). Pupil dilation in response preparation. *International Journal of Psychophysiology*, 67(2), 124-130.

Van Orden K.F. (2001). Eye Activity Correlates of Workload during a Visuospatial Memory Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 43(1), 111-121.*

Veltman, J. A., & Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, *41*(5), 656-669.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 433-441.

Zelinsky, G. J., & Sheinberg, D. L. (1997). Eye movements during parallel–serial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(1), 244.

Zheng B. (2013). Capturing and evaluating blinks from video-based eyetrackers. *Behavior Research Methods,* 45(1), 656-663.