Using Variable-Rate Alerting to Counter Boredom in Human Supervisory Control

Citation: Mkrtchyan, A. A., J. C. Macbeth, E. T. Solovey, J. C. Ryan, and M. L. Cummings. "Using Variable-Rate Alerting to Counter Boredom in Human Supervisory Control." Proceedings of the Human Factors and Ergonomics Society Annual Meeting 56, no. 1 (October 26, 2012): 1441-1445.

As Published: http://dx.doi.org/10.1177/1071181312561406

Publisher: Human Factors and Ergonomics Society

Persistent URL: http://hdl.handle.net/1721.1/81782

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

Using Variable-Rate Alerting to Counter Boredom in Human Supervisory Control

Armen A. Mkrtchyan, Jamie C. Macbeth, Erin T. Solovey, Jason C. Ryan, M. L. Cummings
Humans and Automation Laboratory, Massachusetts Institute of Technology, Cambridge, MA

A low task load, long duration experiment was conducted to evaluate the impact of cyclical attention switching strategies on operator performance in supervisory domains. The impetus for such a study stems from the lack of prior work to improve human-system performance in low task load supervisory domains through the use of design interventions. In this study, a design intervention in the form of auditory alerts is introduced and the effects of the alerts are examined. The test bed consists of a video game-like simulation environment, which allows a single operator the ability to supervise multiple unmanned vehicles. Each participant in the study completed two different four hour sessions, with and without the alerts. The results suggest that the alerts can be useful for operators who are distracted for a considerable amount of time, but that the alerts may not be appropriate for operators who are able to sustain directed attention for prolonged periods.

INTRODUCTION AND BACKGROUND

Ever-increasing levels of automation in the past few decades have proved to be advantageous in improving the reliability and safety of systems, as well as their profitability and productivity. Nonetheless, there are drawbacks associated with such increases. Human factors specialists have widely argued that the more advanced the automation is, the more important the role of the operator becomes in successfully monitoring and supervising the automated system (Bainbridge, 1983). Furthermore, increased automation often lowers operator workload, causing boredom and vigilance decrements (Langan-Fox, Sankey, & Canty, 2008; Thackray, 1980). In the past, numerous studies have been conducted to evaluate the effects of boredom on operator performance. More specifically, a study of Air Traffic Control (ATC) tasks revealed that under low traffic conditions, the percentage of operator error due to judgments in planning increased (Rodgers & Nye, 1993). ATC operators who reported high levels of boredom had slower reaction times and worse performance compared to operators who reported low levels of boredom (Thackray, Powell, Bailey, & Touchstone, 1975).

Boredom is closely related to vigilance, which is defined as "a state of readiness to detect and respond to certain small changes occurring at random time intervals in the environment" (Mackworth, 1957). It has been shown that participants of vigilance experiments often report high levels of boredom (Scerbo, 1998). Some researchers stated that vigilance decrements occur under conditions of low workload, when arousal level is low (Manly, Robertson, Galloway, & Hawkins, 1999; Proctor & Zandt, 2008). However, a recent study showed that vigilance tasks can be demanding (Warm, Parasuraman, & Matthews, 2008). It has also been observed that performance declines during vigilance tasks and varies with signal salience (Temple et al., 2000).

Boredom and vigilance problems can be exacerbated by systems with high levels of automation, which leave human operators unengaged for prolonged periods. Many of these systems can be classified as supervisory control systems, in which "one or more human operators are intermittently programming and continuously receiving information from a computer that itself closes an autonomous loop through artificial effectors and sensors to the controlled process or task environment" (Sheridan, 1992).

There are numerous examples of highly automated supervisory control systems that could lead to boredom and vigilance decrement. One example is the operation of the Predator unmanned aerial vehicle. In an interview, a Predator pilot said, "Highly skilled, highly trained people can only eat so many peanut M&Ms or Doritos or whatnot…There's the 10 percent when it goes hot, when you need to shoot to take out a high-value target. And there's the 90 percent of the time that's sheer boredom – 12 hours sitting on a house trying to stay awake until someone walks out" (Button, 2009). Increased automation also contributed to low vigilance exhibited by the Northwest flight 188 crew that overflew Minneapolis-St. Paul Airport by 150 miles in 2009 (The Federal Aviation Administration, 2009). Nuclear power plant control is yet another domain in which boredom and vigilance problems exacerbated by automated systems are not uncommon (Kaku & Trainer, 1992).

Cyclical Attention Management

A recent study on the effects of prolonged low task load on operator performance confirmed that operators' vigilance is a valid predictor of their performance in the context of controlling multiple unmanned vehicles (UVs) (Hart, 2010). More specifically, operators with low vigilance performed worse than operators with high vigilance. However, the study also revealed that distraction in this low task load supervisory environment was not necessarily detrimental for performance, if managed properly. For example, it was observed that the second-best performer exhibited a cyclical task switching strategy that resulted in performance similar to the best performer. Surprisingly, this participant was distracted about 45% of the time, compared to the 10% distraction level of the best performer.

Based on this previous result, it was hypothesized that a design intervention that prompts participants to switch their attention in a cyclical manner could be effective in improving operator performance in low task load supervisory domains. To evaluate the feasibility of prompting participants to switch attention for potential performance improvement, a long duration low task load experiment was conducted, discussed in the next section.

EXPERIMENTAL SETUP

The simulation test bed used in this experiment, the Onboard Planning System for Unmanned Vehicles Supporting Expeditionary Reconnaissance and Surveillance (OPS-USERS), was inspired by a futuristic UV control paradigm, in which a single operator is responsible for monitoring and controlling multiple UVs (Fisher, 2008; Mkrtchyan, 2011). The OPS-USERS system simulates a search and destroy mission, where UVs are tasked to search an area for targets, then track and eventually destroy them. The control structure is based on a high-level, goal-oriented scheme where operators specify locations on a map where they want the vehicles to search for targets, as opposed to a low level control scheme requiring operators to specify altitude, heading, airspeed or other vehicle-level parameters. While this experiment used the simulation version of this test bed, OPS-USERS can be used to operate multiple actual air and ground unmanned vehicles (Kopeikin, Toupet, Clare, Cummings, & How, 2012).

Hardware

An operator workstation consisted of a Dell Inspiron desktop computer with a 17 inch monitor that was dedicated to running the OPS-USERS interface. A second 17 inch monitor was available for the operators to use for non-simulation related purposes. The operators were videotaped using Microsoft™ HD web cameras for the duration of the experiment. One camera was allocated per operator and another camera recorded the overall view of the experimentation room. Additionally, all participants were required to wear wireless headphones, which allowed them to move around the experimentation room and still be able to hear auditory alerts of the OPS-USERS interface.

Participants

Nine participants were tested in groups of three in order to simulate typical unmanned vehicle operating environments. Each participant worked individually at a workstation running an independent version of OPS-USERS. Participants were compensated $400 for their participation in two four-hour studies, which were administered on different days. In addition, they were informed that the person with the highest performance score would receive a $250 BestBuy gift card. Two females and seven males were recruited from the undergraduate and graduate student population of MIT. Ages ranged from 18 to 24, with Mean (M) of 20.7 years and Standard Deviation (SD) of 1.4 years.

Experimental Procedure

Participants were first asked to complete a demographic survey, indicating their age, gender, occupation, military experience, video gaming experience, sleep duration for the past two nights, and comfort level using computers. The NEO-FFI-3 personality survey, which rates participants' neuroticism, extraversion, openness to new experiences, agreeableness, and conscientiousness (McRae & Costa, 2010), was also administered. Lastly, a Boredom Proneness Survey (BPS) (Farmer & Sundberg, 1986) was administered. All participants then completed a training session consisting of a self-paced PowerPoint™ tutorial and a practice session using the OPS-USERS interface.

During the test session, each participant was responsible for controlling four UVs. Over the course of the four hour test period, six targets were available to be found, half of which were hostile and needed to be destroyed. During the test, participants were allowed to interact with each other and use personal items, such as books, laptops and cell phones for data, but cell phone calls were not permitted. Additionally, snacks and a variety of non-alcoholic beverages were provided. All these items served as possible distractions from the OPS-USERS interface. After the conclusion of the test session, participants completed a post-experiment survey, detailing their confidence level, busyness level, and the usefulness of auditory alerts on a five-point Likert scale.

Experimental Design

The study was conducted to evaluate the effects of cyclical attention switching strategies on operator performance in low task load supervisory domains. For this reason, each participant completed two four-hour test sessions: one with a design intervention to prompt cyclical attention switching and another test session without the design intervention. The order of the sessions was randomized and counterbalanced to avoid carryover effects. The intervention was implemented in the form of auditory alerts that were pre-programmed in the interface. The alerts consisted of four distinct chimes approximately 300ms long that resembled a doorbell sound. Between the first two and last two chimes there was a 400 ms pause. Between the second and the third chimes the duration of the pause was approximately 1.2 seconds. All participants wore the required wireless headphones at all times to hear the alerts. The number of the alerts changed in a cyclical pattern, which can be described by Equation 1. Figure 1 shows the number of alerts across four hours.

The independent variable in this experiment was the presence of the alerts. The dependent variables were utilization, performance scores, participants' attention states, and subjective, self-rated metrics.

$$f(t) = 40 + \left(40 - \frac{t}{24}\right) * \cos\left(\frac{t}{14 - \frac{t}{120}}\right), t \in [0,240] \quad (1)$$

where t is the time of the experiment in minutes.

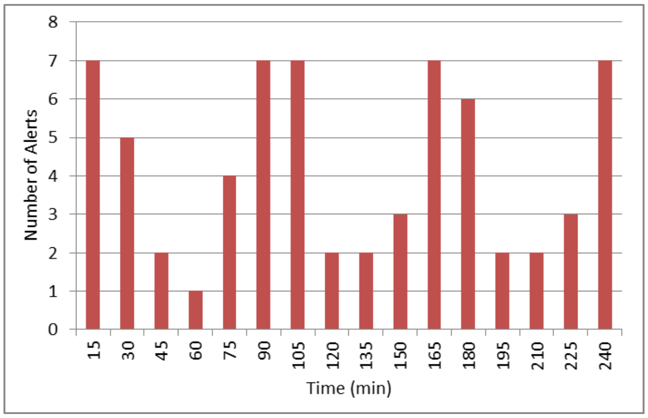

Figure 1: Histogram showing the cyclical pattern of alerts over time

Measures

As a measure of objective workload, utilization was used. It is defined as the "percent busy time" – the time operators spent performing various tasks in the interface divided by the total available time. Although utilization does not account for the time that operators monitor the simulation, it is a useful metric that measures operator interaction with a system and has been used to detect changes in workload (Cummings & Guerlain, 2007).

Two different performance scores, the Target Finding Score (TFS) and the Hostile Destruction Score (HDS), provide information on how well the objectives of the mission were accomplished (Mkrtchyan, 2011). The TFS accounts for the speed of finding targets and quantity of targets found. The HDS accounts for the speed of destroying hostile targets and quantity of destroyed targets. The performance score ranges from zero to two, where a higher score is better.

Operators' attention states were estimated by classifying their video-taped activities. Three categories of attention states were identified: directed, divided, and distracted. In the directed attention state, the operator monitors or interacts with the simulation interface. In the divided attention state, the operator monitors the interface while multitasking (i.e., eating while monitoring the interface). Lastly, in the distracted attention state, the operator is not paying attention to the interface at all. For this state, operators were coded as distracted if they were not in a physical position to see the interface, i.e., turned around in their chair or working on a personal laptop.

RESULTS AND DISCUSSION

An alpha of 0.05 was used for all statistical tests.

Utilization

The required average utilization in the study was 2.1%, based on the number of tasks that operators were required to complete over the course of the study. The total utilization was based on the total number of tasks completed by the operators, including unrequired tasks that operators inserted into the system such as changing a search area. Total utilization was significantly greater than the required utilization. The average total utilization was 14.6%, approximately seven times greater than the required utilization. This was mainly due to the fact that participants of the study interacted with the interface much more than the system required.

The results also indicate that the design intervention did not affect the workload of the operators. A within subject t-test confirmed that there is no statistical difference between the utilization of the first and second sessions ($t(8) = 1.72, p = 0.13$) and between the utilization of the two alerting scenarios ($t(8) = 0.06, p = 0.95$).

Performance Scores

According to a within-subjects t-test, there is no statistical difference across the sum of the performance scores of the two sessions ($t(8) = -1.33, p = 0.22$) and the two scenarios ($t(8) = 1.87, p = 0.1$).

Unexpectedly, the alerts seemed to negatively impact participants' performance scores (Figure 2). As detailed in the next section, this is most likely due to the fact that most of the operators directed their attention to the interface for a majority of the time, thus the auditory alerts were not as necessary in prompting them to pay attention as expected. Also, participants mentioned that they were sometimes confused by the alerts, because they could not understand why they were being prompted to pay attention to the interface when they had already been interacting with the interface for some time.

Attention States

In order to evaluate the attention states, two researchers watched the recorded videos and coded participants' attention states. Eighteen four-hour-long videos were coded (two four-hour videos per participant). Although participants were mostly directed (M 64%, SD 15%) and distracted very little (M 12%, SD 8%), they became less directed during the second session ($t(8) = -4.33, p = 0.01$). More specifically, participants spent on average 58% (SD 8%) in a directed attention state, 27% (SD 5%) in a divided attention state, and 15% (SD 6%) in a distracted attention state. As several participants mentioned in a post-experiment interview, after the first session they became more familiar with the interface and did not have to spend as much time monitoring the system to feel satisfied that they were achieving the objectives of the mission.

Figure 3 shows the allocation of attention states across the two alerting conditions. Nearly equal proportions of attention

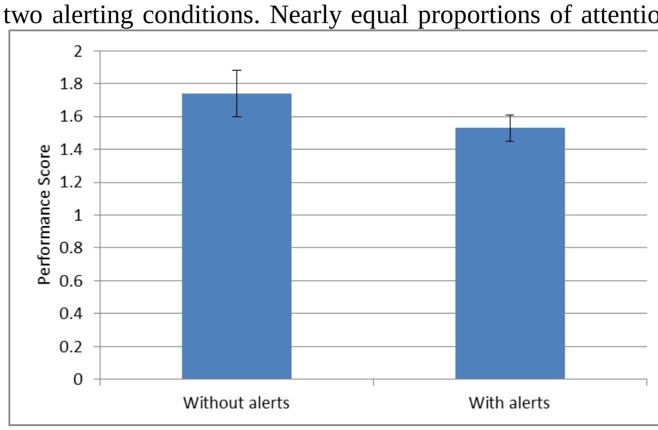

Figure 2: Performance scores

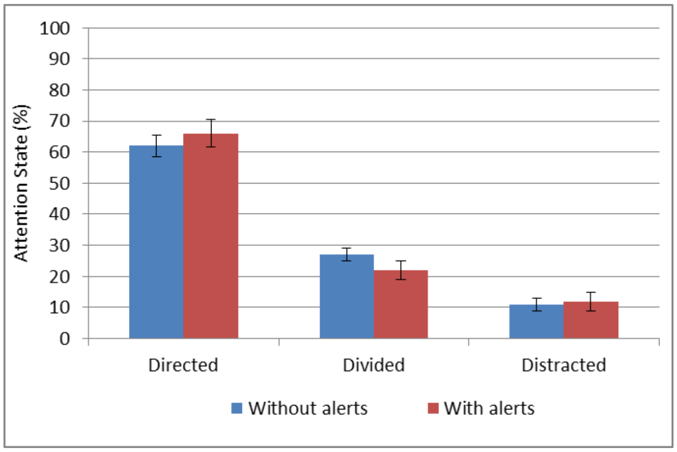

Figure 3: Attention state comparisons across the two alerting conditions

states across the two scenarios (with and without the alerts) indicate that the alerts did not significantly affect the overall allocation of participants' attention resources. A paired t-test confirmed that no statistical difference exists ($t(8) = 0.71, p = 0.49$).

Overall, attention state results differed greatly from an earlier low task load study (Hart, 2010). In the prior study, experiment participants were directed 34% (SD 15%) and distracted about 44% (SD 20%) of their time, while in this experiment participants were highly directed (64%) and minimally distracted (12%).

When comparing the combined directed and divided attention states and the number of alerts across all participants, there was a significant positive correlation for only one of the participants (Spearman's $\rho = 0.87, p < 0.01$). This participant was the least directed among all the participants in the scenario with the alerts. He was directed on average 40% of the time, divided 29%, and distracted 31% of the time.

Figure 4 shows the combined percentage of directed and divided attention states of this participant for the two scenarios. The figure also depicts the number of auditory alerts over time. Across the two scenarios, this participant's performance score was lower during the scenario with the design intervention, which was his first session. However, compared to the mean performance of all participants in each alerting scenario, this participant scored higher than the mean for participants with the alerts and lower on the session without the alerts. Therefore, the design intervention seemed to work for the most distracted participant, leading him to switch his attention in a more cyclical pattern and improving his performance as compared to the average.

The design intervention appeared to work for this participant because he was not directed as much as the rest of the participants and the alerts appeared to prompt him to pay attention to the system. Also, this participant's attention allocation was the most comparable to the attention allocation of the participant in a previous study (Hart, 2010) after whom the cyclical alert system was modeled. More specifically, attention states of this participant (directed 37%, distracted 45%, divided 18%) were closely matched by the attention states of the least directed participant.

Subjective Metrics

Self-rated Metrics. Participants' self-rated metrics provide valuable subjective information on their perceived performance during the experiment. Various subjective metrics, such as self-rated confidence and performance, busyness, and usefulness of alerts were assessed on a five-point Likert scale. Generally, participants indicated low busyness levels, and high self-rated performance. Across the two scenarios, only self-rated confidence was marginally significant ($Z = -1.73, p = 0.083$).

Personality Inventory and BPS Scores. To evaluate whether the personality dimensions were correlated with performance scores, Spearman's correlation test was used. There were no strong correlations between the personality dimensions and performance scores of the two scenarios. However, conscientiousness was marginally correlated with operator performance scores in the scenario without the alerts (Spearman's $\rho = 0.65, p = 0.06$). Interestingly, the mean conscientiousness score for participants was lower than the average for the US population (although not statistically significant).

Lastly, the 28-item BPS was used to assess participants' boredom proneness levels. According to previously conducted studies (Winter, 2002), the sample mean of the US population is around 10.5. Participants who score below 5 are very low on the BPS and those who score above 15 are very high. The results revealed that the majority of the participants had low boredom proneness levels. More specifically, the average BPS score was 7.8 (SD 4.0), minimum score was 4.0 and maximum score was 16.0 on a 28-point scale. Given the low BPS scores, it is not surprising that, on average, participants were only 12% distracted during the experiment.

To assess whether the BPS score could be used to predict operator performance, correlation coefficients between the BPS scores and the performance scores were calculated. The results indicate that no significant correlation exists in this data set. This is important, since it suggests that boredom proneness was not a major factor affecting participants' performance. In fact, the best and the worst performers exhibited the same level of boredom proneness.

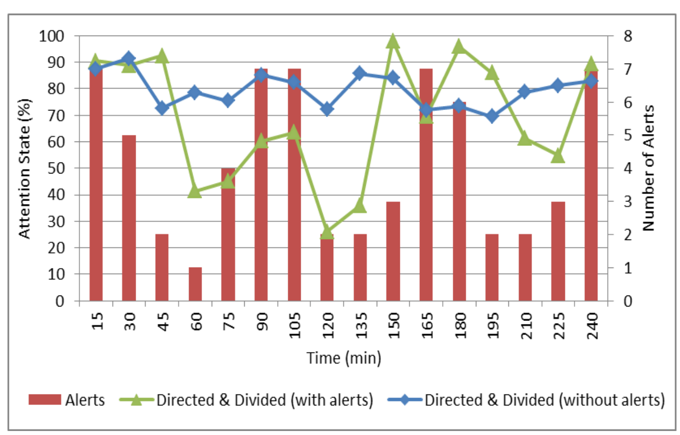

Figure 4: Attention state comparisons across the two scenarios for the least directed participant

CONCLUSIONS AND FUTURE WORK

This paper presents a study that aimed to evaluate the effectiveness of cyclical attention switching strategies in low task load supervisory domains. To prompt the participants of the study to switch their attention in a cyclical manner, auditory alerts were utilized, where the number of these alerts was varied over time in a sinusoidal pattern. Results of the study indicate that participants were significantly different from a previous population in terms of their ability to sustain attention for prolonged periods. On average, participants had a low propensity of being bored. Over the course of the study, the participants were distracted only about 12% of the time, which is remarkable given the very low task load nature of the experiment.

Objective workload measured through utilization indicates that participants interacted with the interface significantly more than required, and most of the participants performed much better compared to a previously conducted, similar experiment.

The design intervention implemented in the experiment to help operators of supervisory systems sustain directed attention could not be validated to have positive effects. This is most likely due to the fact that the participants, in general, were highly directed. However, it should be mentioned that the participant who was the most distracted exhibited a cyclical attention switching strategy in the scenario with the design intervention. Moreover, this participant, despite being the most distracted, performed better than average, indicating that the design intervention can be useful for more distracted participants.

In the future, to fully evaluate the design intervention, a new low task load, long duration study needs to be conducted with a new set of participants who have difficulties sustaining directed attention. Thus, a better selection process for participants needs to be developed to select participants who have difficulties sustaining attention over prolonged periods of time. This is the subject of current research.

Also, the auditory alerts that were implemented in the experiment were set *a priori* and did not rely on operator performance or on parameters of the mission. Further analysis should be conducted to determine whether it is more appropriate to have auditory alerts based on operator interaction patterns, mission tasks, or other parameters that might help identify the “right” time for the intervention.

ACKNOWLEDGMENTS

This research was supported by the Office of Naval Research (ONR). The authors would like to thank Kristopher Thornburg, Andrew Clare, and Lucy Du for their help in conducting the experiment and collecting results.

REFERENCES

Bainbridge, L. (1983). Ironies of Automation. *Automatica, 19*, 775-779.

Button, K. (2009). Different Courses: New Style UAV Trainees Edge Toward Combat. *Training and Simulation Journal*, 42-44.

Cummings, M. L., & Guerlain, S. (2007). Developing Operator Capacity Estimates for Supervisory Control of Autonomous Vehicles. *Human Factors, 49*(1), 1-15.

Farmer, R., & Sundberg, N. D. (1986). Boredom Proneness - The Development and Correlates of a New Scale. *Journal of Personality Assessment, 50*(1), 4-17.

Fisher, S. (2008). *Replan Understanding for Heterogenous Unmanned Vehicle Teams.* Massachusetts Institute of Technology, Cambridge, MA.

Hart, C. S. (2010). *Assessing the Impact of Low Workload in Supervisory Control of Networked Unmanned Vehicles.* Massachusetts Institute of Technology, Cambridge, MA.

Kaku, M., & Trainer, J. (1992). *Nuclear Power: Both Sides: The Best Arguments For and Against the Most Controversial Technology*. New York: W. W. Norton & Company, Inc.

Kopeikin, A., Toupet, O., Clare, A. S., Cummings, M. L., & How, J. P. (2012). *Outdoor Flight Tests of a Decentralized Multi-UAV System with Human Supervision*. Paper presented at the AIAA Guidance, Control, and Navigation Conference.

Langan-Fox, J., Sankey, M. J., & Canty, J. M. (2008). Keeping the Human in the Loop: From ATCOs to ATMs, *Keynote Speech by J. Langan-Fox at the Smart Decision Making for Clean Skies Conference*. Canberra, Australia.

Mackworth, N. H. (1957). Some Factors Affecting Vigilance. *Advancement of Science, 53*, 389-393.

Manly, T., Robertson, I. H., Galloway, M., & Hawkins, K. (1999). The Absent Mind: Further Investigations of Sustained Attention to Response. *Neuropsychologia, 37*, 661-670.

McRae, R. R., & Costa, P. T. (2010). *NEO^{TM} Inventories*. Lutz, FL: PAR Publishing.

Mkrtchyan, A. (2011). *Modeling Operator Performance in Low Task Load Supervisory Domains.* MIT, Cambridge, MA.

Proctor, R. W., & Zandt, T. V. (2008). *Human Factors in Simple and Complex Systems* (2nd ed.). Boca Raton, FL: CRC Press.

Rodgers, M. D., & Nye, L. G. (1993). Factors Associated with Severity of Operational Errors at Air Route Traffic Control Centers. In M. D. Rodgers (Ed.), *An Examination of the Operational Error Database for Air Traffic Control Centers* (pp. 243-256). Washington D.C.: Federal Aviation Administration, Office of Aviation Medicine.

Scerbo, M. W. (1998). What's So Boring About Vigilance? In R. R. Hoffman, M. F. Sherrick & J. S. Warm (Eds.), *Viewing Psychology as a Whole: The Integrative Science of William N. Dember* (pp. 145-166): American Psychological Association.

Sheridan, T. (1992). *Telerobotics, Automation and Human Supervisory Control*. Cambridge, MA: The MIT Press.

Temple, J. G., Warm, J. S., Dember, W. N., Jones, K. S., LaGrange, C. M., & Matthews, G. (2000). The Effects of Signal Salience and Caffeine on Performance, Workload, and Stress in an Abbreviated Vigilance Task. *Human Factors, 42*(2), 183-194.

Thackray, R. I. (1980). Boredom and Monotony as a Consequence of Automation: A Consideration of the Evidence Relating Boredom and Monotony to Stress, *DOT/FAA/AM-80/1*. Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

Thackray, R. I., Powell, J., Bailey, M. S., & Touchstone, R. M. (1975). Physiological, Subjective, and Performance Correlates of Reported Boredom and Monotony While Performing a Simulated Radar Control Task. Oklahoma City: Federal Aviation Administration, Civil Aeromedical Institute.

The Federal Aviation Administration. (2009). Northwest Airlines Flight 188. Retrieved January 12, 2011, from http://www.faa.gov/data_research/accident_incident/2009-10-23/

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance Requires Hard Mental Work and is Stressful. *Human Factors, 50*(3), 433-441.

Winter, R. (2002). *Still Bored in a Culture of Entertainment: Rediscovering Passion & Wonder*. Nottingham, UK: IVP Books Publishing.