Circadian Effects on Simple Components of Complex Task Performance

Benjamin A. Clegg[1], Christopher D. Wickens[1,2], Alex Z. Vieane[1], Robert S. Gutzwiller[3], & Angelia L. Sebok[2]
[1]*Colorado State University*, [2]*Alion Science and Technology*,
[3] *Space and Naval Warfare Systems Center Pacific*

The goal of this study was to advance understanding and prediction of the impact of circadian rhythm on aspects of complex task performance during unexpected automation failures, and subsequent fault management. Participants trained on two tasks: a process control simulation, featuring automated support; and a multi-tasking platform. Participants then completed one task in a very early morning (circadian night) session, and the other during a late afternoon (circadian day) session. Small effects of time of day were seen on simple components of task performance, but impacts on more demanding components, such as those that occur following an automation failure, were muted relative to previous studies where circadian rhythm was compounded with sleep deprivation and fatigue. Circadian low participants engaged in compensatory strategies, rather than passively monitoring the automation. The findings and implications are discussed in the context of a model that includes the effects of sleep and fatigue factors.

INTRODUCTION

With an ever-increasing proliferation of systems featuring automation, understanding and predicting operator performance when automation unexpectedly fails has naturally become a critical issue. Developing a detailed understanding is of value to domains in which errors can have catastrophic consequences, wherein an individual may need to operate with little or no support, and where a range of factors may degrade the expected level of human performance. The CODDMAN (Complacency, Detection, Diagnosis, and Fault Management) model represents an attempt to predict how a variety of factors affect the performance of an operator during a sudden workload transition, of the type associated with an unexpected failure of an automated system within these environments. Our interest is particularly in the environment of the astronaut on long duration space missions (Sebok, Wickens, Clegg, & Sargent, 2014), however the same factors will be relevant for a variety of scenarios featuring automation failures.

To account for the effect of sleep disruption on the complex and multi-tasking performance required for unexpected failure management within the model, an initial review and meta-analysis on sleep disruption effects on complex task performance (Wickens, Hutchins, Laux, & Sebok, 2015) was conducted. Relevant to the current research, two important findings emerged: (1), sleep-fatigue-induced decrements on complex cognitive and multi-task performance were considerably less severe than those reported from simpler tasks which involved reaction time and vigilance. In addition (2), circadian night created a threefold magnification of performance decrements associated with sleep deprivation, compared to circadian day.

Incorporating such factors into the CODDMAN model enables predictions about the effects of fatigue and sleep disruption on each stage of automation fault management (see Sebok et al., 2015, this symposium). Single task monitoring and vigilance component tasks preceding the unexpected failure are likely to be more effected by sleep deprivation than the more complex components of the diagnosis and fault management phases of CODDMAN (see Figure 1).

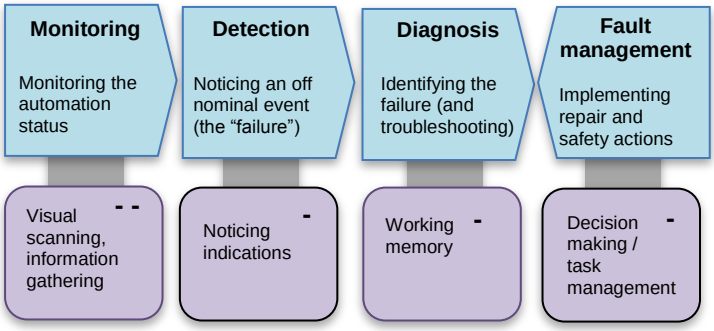

Figure 1. The Fatigue and Sleep factors from the CODDMAN model of operator performance. Minus signs indicate factors associated with impoverished performance, with double signs showing greater magnitude effects.

Reacting appropriately to the onset of a fault depends on prior monitoring and developing an understanding of the system state. Comprehension of the nature of the evidence available is a vital component of fault detection that might be masked by failing to notice that an event has occurred. As described below, this observation led us to operationalize monitoring and detection in a different fashion than is typical elsewhere in the literature on fatigue effects. One specific aspect was the inclusion in our experiments of a salient master alarm, which always alerted participants to the onset of a fault, thus eliminating fatigue effects associated with a simple failure of visual search.

One set of predictions about variations in performance from circadian rhythm effects comes from a study by Manzey and colleagues (Manzey, Reichenbach, & Onnasch, 2009) who observed that fatigued operators tended to compensate for sleep-related fatigue through increased sampling of information. A tendency to shift from passive supervisory observation of automation, to more active and

engaged behavior, would be consistent with a compensatory reaction against the taxing demands found in monitoring an automatic system (Warm, Parasuraman, & Matthews, 2008). Thus, an adaptive reaction to the high demands of passive visual monitoring might be to transform the activity as an operator to more *active* inspection of the system state.

Time of day effects within two complex tasks

The current experiment employed two different types of tasks to examine fatigue effects on simple versus complex aspects of task performance. A supervisory process control task included operations from all 4 stages of the CODDMAN model; and a multi-tasking platform included one specific routine monitoring element amid higher demand tasks, which also included switching between tasks. These multi-tasking requirements are an important domain for the model because it is prototypical of the kinds of demands placed on operators in failure management.

In the supervisory process control task AutoCAMS (Manzey et al., 2008), which is designed to simulate environmental control in space, operators attempt to keep Oxygen and Nitrogen levels within a safe range. Further, operators must identify, diagnose, and repair failures as they occur during times when reliable automated assistance is available (Routine Failure), when automated assistance is available but incorrect (First Failure of Automation), and when automated assistance is not available (Second Failure of Automation).

For assessing multitasking performance, we used MATB (Multi Attribute Task Battery) II (Santiago-Espada, Myer, Latorella, & Comstock, 2011). This is a multi-tasking research platform that requires individuals to oversee four concurrent subtasks (tracking, monitoring, resource management, and communications). All of the subtasks were presented visually on a screen in four quadrants so that each task was visible at all times, except for the information used in the communications task. The communications task required participants to listen to a simulated air traffic control message and respond to specific messages if they were directed at the participant's identification number. Participants were, by a single-handed control, able to perform only a single subtask at a time.

The goal of the current study was to identify the effect of time of day (circadian day versus night) on CODDMAN components. Consistent with the review above, three major questions were addressed:

1) Are the effects of sleep related-fatigue on multi-tasking and complex diagnosis less than those observed on vigilance or monitoring?

2) Are the same active-engagement compensatory effects of circadian-induced fatigue found by Manzey et al. (2009) observed without the sleep loss component?

3) Within AutoCAMS, if effects due to visual scanning components are mitigated through the use of a master alarm, do time-of-day effects still occur?

METHODS

Participants

Fifty-six participants signed up for three separate sessions, and received $45 compensation for their attendance. Of those, 50 attended an AutoCAMS session of the experiment but 1 participant failed to understand the task leaving 49 participants for that task, and 47 students completed the MATB portion.

Materials & Procedure

Participants were trained midday (between 10 am – 2 pm) on both AutoCAMS 2.0 and MATB II. Training sessions were designed to last 90 minutes, where 30 minutes was allocated to learn MATB and 60 minutes allocated to learn AutoCAMS 2.0. Training on the supervisory process control task (AutoCAMS) consisted of a self-paced multimedia presentation in PowerPoint. Participants were introduced to the simulation of the life-support system, where they had to maintain Oxygen and Nitrogen within normal range to ensure safe conditions for a crew of astronauts. Once a failure was introduced into the system, which caused levels to go out of target range, operators had to detect, diagnose, repair, and manually manage the system to return levels to normal. AutoCAMS training took approximately 30-40 minutes to complete. After training, participants completed a 5 minute practice block in AutoCAMS where an automated decision aid, AFIRA, correctly identified a failure and provided steps to take to manage the failure. One routine failure occurred during the practice block. Once the practice block ended, participants were able to ask questions to clear up any misunderstanding. MATB training consisted of a series of slides in PowerPoint that was adapted from Santiago et al. (2011). MATB training was self-paced, and took approximately 15 minutes to complete plus a brief two-minute practice trial in which all four tasks presented task events. Similarly, participants were able to ask questions about the task during this period.

To induce circadian effects without sleep deprivation (i.e., time of day effects without requiring participants to remain awake into the night), test phase sessions were conducted at either 5am ("circadian low", and still part of the circadian night phase identified by Wickens et al., 2015), or at 5pm ("circadian high", and still part of the circadian day phase).

Participants were therefore required to come back 2-3 days after their training session for either a 5am session or a 5pm session (retention delay between training and test phases varied between 29 and 67 hours; M= 51.8 hours, SD= 11.7). Those who had a 5am session first returned that same day for a 5pm session. Those whose first experimental session was at 5pm came back the following day at 5am for a second session. The order of sessions, and the task performed in them (AutoCAMS or MATB) was counter-balanced.

AutoCAMS. A 2 (Time of day: circadian low vs. circadian high) X 3 (Failure type: Automation support with a decision aid ("**Auto**"), "First Failure" (**FF**) of automated support,

"Second failure" (**2F**)) MANOVA was used with fatigue condition as a between subjects variable, and failure type as a within subjects variable. Consistent with the method from Wickens, Clegg, Vieane, and Sebok (2015) participants completed four blocks of trials, each featuring system faults but with different levels of automation aid. All participants were provided with a master alarm in all blocks to alert the onset of failures, assisting with the detection portion of the task. Within the first two blocks, routine failures were coupled with correct AFIRA diagnosis and management guidance. In block 3, the automated assistance failed (FF) by providing incorrect information (a wrong diagnosis and incorrect management steps). Block 3 lasted 15 minutes, with the fault (and associated automated aid failure) introduced 10 minutes into the block. In block 4, a fault again occurred, but the automated assistance was made unexpectedly unavailable for participants, leaving them to diagnose, repair, and manually control the system on their own. In block 4 the failure (2F) occurred 1 minute into the 5 minute block.

When a failure occurred, as indicated by the master alarm changing from green to red, participants were expected to diagnose the failure, initiate a repair order, and then manually manage the failing system to return levels to within their target range. In training, participants were informed that the automation could potentially fail, and of the importance of verifying the diagnosis provided by the automation. The automated assistance provided both the diagnosis of the failure (e.g., "Oxygen valve leak") and steps to manage that failure (e.g., "Turn Oxygen flow to high"). When automated assistance was absent (the 2F event), participants had to rely on their own abilities to diagnose, repair, and manage the failure. Monitoring, detection, diagnosis, and management performance was collected for all experimental trials.

MATB. In the MATB task, participants completed three test trials each ten minutes in length, comprising an easy, difficult, and mixed tracking difficulty condition (the entire procedure duplicated Gutzwiller et al., 2014).

The tracking task was present for the entire duration of all trials. Trials consisted of equal numbers of competing events. In the communications task, these comprised both own-ship, and other-ship auditory instructions. Participants only needed to respond to own-ship events. In the monitoring task, participants were asked to monitor for the onset of a red light, the offset of a green light, and for indicators on the four scale measures to "stick" in the upper or lower region. All of these events required a click before a 10-second timeout in response to reset the indicator. Events in the resource management task were failures of the eight different pumps that regulate the flow of fuel to two main, constantly depleting tanks. When a pump fails, participants had to route the flow through activating or deactivating other pumps until the failed pump reset after 30s.

Measures

Performance on the CODDMAN components with AutoCAMS was operationalized using the following metrics:

Monitoring: The average number of inspections per minute was used to determine how often an individual was checking both Oxygen and Nitrogen levels to ensure that they were within range.

Detection: A visually salient alarm (changes from green to red) was used to notify the participant that a failure was present. When using such an "alarm" we measured the time to detect a failure as the difference in time between the occurrence of the failure (e.g., alarm changes from green to red) and the time of the first diagnosing action in AutoCAMS. Typically, poor monitoring behavior is related to poor ability to detect failures (e.g., Metzger & Parasuraman, 2005). However, in this experiment, we anticipated that these two measures would be independent, as the alarm indicated a failure without necessitating any monitoring clicks.

Diagnosis: Diagnostic accuracy was calculated as the ratio of the number of correct repairs to the number of total repairs. The completeness of the diagnostic process reflects the number of steps (out of 5) taken to be sure of any one diagnosis. A dichotomous measure was also used to identify whether a participant was able to successfully repair the failure in the time allotted.

Fault Management: The fault management success of the participant in keeping levels within range was measured by the time (in seconds) that the failing system (Oxygen or Nitrogen) spent outside of the target range.

RESULTS

AutoCAMS

The data presented includes data from the third routine failure that participants encountered which included automated decision support ("**Auto**"). Data from the "first failure" (**FF**) of AFIRA decision support in block 3, and the "second failure" (**2F**) of AFIRA in block 4, are also presented.

Monitoring. Frequency of monitoring clicks was collected from before the failure was injected (pre-failure), as shown in Figure 2.

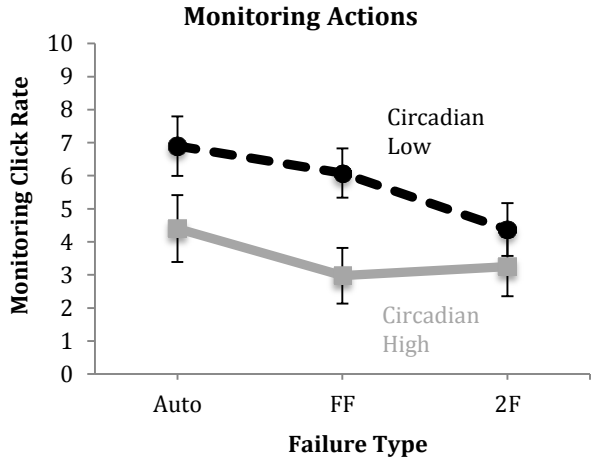

Figure 2. Monitoring behavior across failure types by fatigue condition. Black dashed line shows circadian low, solid grey line circadian high participants. (Error bars show standard error).

There was a significant main effect of time of day on monitoring behavior ($F(1,41)= 4.50$, $p< .05$) where participants performing during circadian low monitored the Oxygen and Nitrogen systems more than those who were circadian high, replicating the trend found by Manzey et al. (2009). There was a significant main effect of the failure type (Wilks' $\lambda = .84$, $F(2,40)= 3.89$, $p< .05$) consistent with less monitoring over time. The failure type by time of day interaction was marginally non-significant (Wilks' $\lambda = .89$, $F(2,40)= 2.60$, $p= .09$). Importantly, in the case when the automation failed for the first time (**FF**) providing incorrect advice, those participants who were circadian low initiated more monitoring clicks per minute, compared to the circadian high group ($t(47)= 2.28$, $p< .05$, $d= .67$). Once the decision support system had failed once, both groups showed similar monitoring behavior ($t(43)= 0.94$, $p> .05$, $d= .29$), within the trial prior to the second failure (**2F**).

Detection. Consistent with the presence of the master alarm to aid detection, there was no main effect of time of day condition ($F(1,39)<1$), with times to the first action the same for the circadian low participants ($M= 6.71$, $SE= .79$) compared to the circadian high participants ($M= 6.79$, $SE= .87$). There was no main effect of failure type (Wilks' $\lambda = .92$, $F(2,40)= 1.71$, $> .05$), and no significant time of day by failure type interaction (Wilks' $\lambda = .94$, $F(2,39)= 1.18$, $p> .05$).

Diagnosis.
Number of diagnostic steps taken. For the diagnostic steps undertaken, there was a main effect of failure type, with increased elements of diagnosis carried out prior to a repair in the 2F condition (Wilks' $\lambda = .27$, $F(2,41)= 56.62$, $p< .01$). There was no significant main effect of time of day ($F(1,42)= 1.09$, $p> .05$). There was a marginally non-significant time of day by failure type interaction (Wilks' $\lambda = .88$, $F(2,41)= 2.73$, $p= .08$). The two groups were similar on the AFIRA-supported failure and the second failure. However, on the first failure, the low circadian group actually performed **better,** in terms of their diagnostic thoroughness, reacting by increasing their diagnostic effort when presented with an unsignaled automation failure with incorrect advice compared to the previous correct automation advice ($t(24)= 2.45$, $p< .05$). In contrast the high circadian group showed complacency in treating the incorrect advice the same as correct advice ($t(21)= 0.00$, $p> .05$). Such an advantage for the circadian low group is plausibly related to their more vigilant monitoring behavior (see Figure 3).

Diagnostic accuracy. There was a significant main effect of failure type (Wilks' $\lambda = .31$, $F(2,41) = 44.69$, $p< .01$), where participants' accuracy dipped on the unexpected first failure of the decision support system, and remained somewhat lower with the support (**Auto**: $M= .89$, $SE= .04$; **FF**: $M= .17$, $SE= .05$; **2F**: $M= .66$, $SE= .04$). There was no significant main effect of time of day condition ($F(1,41)= 1.33$, $p> .05$) and no significant failure type by time of day interaction (Wilks' $\lambda = .998$, $F(2,41)< 1$).

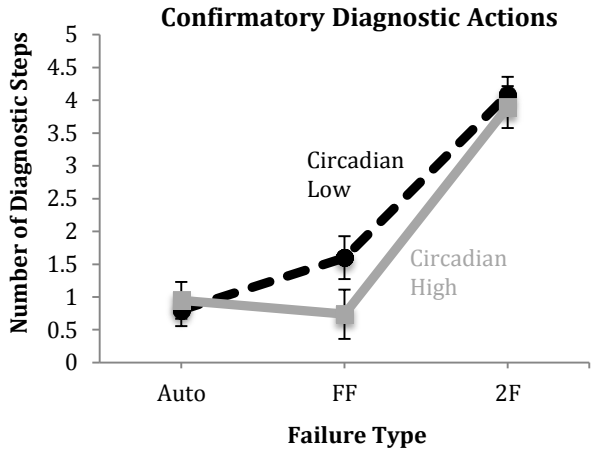

Figure 3. Confirmatory diagnostic steps prior to first repair attempt across failures types by fatigue condition. Black dashed line shows circadian low, solid grey line circadian high participants. (Error bars show standard error.)

Failure Management. There was a significant main effect of failure type (Wilks' $\lambda = .58$, $F(2,41)= 15.08$, $p< .01$), where the unexpected first failure of the decision support system left the affected system out of range longer, and also proved harder to maintain during the second automation failure (**Auto**: $M= 84.9$s, $SE =13.9$; **FF**: $M= 177.4$s, $SE= 11.9$; **2F**: $M= 130.2$s, $SE= 7.3$). There was no significant main effect of time of day condition ($F(1,42)< 1$) and there was no significant failure type by time of day interaction (Wilks' $\lambda = .98$, $F(2,41)< 1$).

Multi Attribute Task Battery

Performance in MATB was separated into that of four main tasks; responding to the communications events quickly and accurately, keeping error in the resource management and tracking tasks low, and responding quickly and accurately to the monitoring task events. No differences in performance were found between circadian high and low for reaction time to communications events ($F<1$), their accuracy ($F(1,42)= 1.67$, $p> .05$), tracking error ($F<1$) or a log transform of resource management error ($F(1,42)= 3.01$, $p= .09$). The marginally non-significant effect in resource management showed the circadian low group had slightly less error ($M= 2.46$, $SE= .09$) than the circadian high group ($M= 2.66$, $SE= .07$).

Only monitoring evinced a significant reduction in performance under circadian low overall. Of the three events to be monitored, a scale deflection, a green light offset and a red light onset, the latter of these showed a large and highly significant decrement in detection rate, from 85% to 64% in the circadian low condition, relative to the circadian high condition after correcting for violations of Levene's test ($t(35)= 2.56$, $p< .05$). This decrement only appeared when the concurrent tracking task was at its difficult level, and hence the monitoring task, generally rated as lowest priority (Gutzwiller et al., 2014) received the fewest resources. Additionally, there was no impact of circadian time on overall switching frequency ($t(42)= -1.43$, $p= .16$).

DISCUSSION

The present experiment examined three hypotheses regarding the effect of circadian-induced fatigue on performance. This fatigue manipulation was employed as a proxy for other variables – total sleep disruption and multiple nights sleep restriction – which had been observed in our meta-analyses to produce qualitatively equivalent effects (Wickens, Hutchins et al., 2015). Within the constraints of our current population, the circadian manipulation proved to be the most ready way of inducing sleep disruption in a controlled manner to examine the predictions of, and extend, the CODDMAN model.

First, based on the findings of the meta-analyses, in contrast with other findings in the fatigue literature, we hypothesized that the effects would be more pronounced in simple, rather than in more complex tasks. Across all tasks, from both platforms (AutoCAMS and MATB) a differential trend was observed here, consistent with the hypothesis. That is, the complex task components of diagnosis and management (in AutoCAMS) and the multi-tasking components (in MATB) were, with one exception, not degraded at all by circadian induced fatigue. Thus, we argue that the complexity of these tasks, and their general interest and engagement, was sufficient to mobilize compensatory arousal and counteract any possible fatigue-related decrements. Furthermore, with a reasonable sample size in the current study, our statistical power available to detect large effects of the sort to have practical impact on real-world task performance ought to have been sufficient.

The notable exception was red-light monitoring in MATB, precisely the kind of task found to be most disrupted by fatigue (Lim & Dinges, 2010). Subjective ratings had shown this task to be both boring, and of low subjective priority (Gutzwiller et al., 2014).

Our other two hypotheses addressed tasks that might otherwise have been anticipated, in the context of the CODDMAN model, to have been disrupted by fatigue: pre-failure monitoring and failure detection in AutoCAMS. Regarding pre-failure monitoring, we observed the same active engagement compensatory effects of circadian-induced fatigue that Manzey et al. (2009) had observed, under sleep deprivation. Thus, the data suggest that when individuals are at low points in their circadian rhythm, they engage in more active monitoring of the system, an engagement that actually served them well in their diagnosis on the first failure without decision support. Such a finding implies that systems that allow operators or automation to adapt the supervisory role to reflect the circadian cycle might be beneficial.

Regarding <u>detection</u> of the AutoCAMS failures in stage 2 of the CODDMAN model, in the current experiment, we did not find differences in the time to detect a failure. Although we identified detection as a simple task where one might expect to see a difference, any deficit was likely mitigated by the visually salient alarm in combination with the AFIRA dialogue box, which also indicated the presence of an error. With these two attention-grabbing features present at the time of the failure, the presence of a failure was noticeable even in a state of fatigue.

Thus, while generally confirming the hypotheses, our experimental results were slightly surprising in revealing the near complete absence of fatigue-related **decrements** in all aspects of complex performance. The meta-analysis after all, had not shown such decrements to be *absent*; only of smaller magnitude than for simpler vigilance, detection, and reaction-time tasks. In response, we can only infer that our imposition of circadian fatigue turned out to be a less-powerful stressor than we had anticipated, in the absence of accompanying sleep deprivation which has been found to amplify circadian night effects (see Wickens, Hutchins et al., 2015).

ACKNOWLEDGMENTS

This work was supported by NASA under Grant NNX12AE69G (PI: Angelia Sebok), technical monitor Dr. Jessica Marquez and technical sponsor Dr. Brian Gore. RSG's contribution was also supported by a DoD SMART scholarship through Space and Naval Warfare Systems Center Pacific. We thank Tyler Scott for aid in extracting data from MATB.

The views and opinions expressed here are those of the authors and do not necessarily reflect the official policy or position of any agency of the U.S. government.

REFERENCES

Clegg, B. A., Vieane, A. Z., Wickens, C. D., Gutzwiller, R. S., & Sebok, A. L. (2014). The effects of automation-induced complacency on fault diagnosis and management performance in process control. In *Proceedings of the Human Factors and Ergonomics Society Annual, 58*(1), 844-848.

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2014). Workload overload modeling: An experiment with MATB II to inform a computational model of task management. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58*(1), 849-853).

Lim, J., & Dinges, D. F. (2010). A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. *Psychological Bulletin, 136,* 375–389.

Manzey, D., Bleil, M., Bahner-Heyne, J.E., Klostermann, A., Onnasch, L., Reichenbach, J., & Röttger, S. (2008). AutoCAMS 2.0 manual. Retrieved from http://www.aio.tu-berlin.de/?id=30492

Manzey, D., Reichenbach, J., & Onnasch, L. (2009). Human performance consequences of automated decision aids in states of fatigue. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 53*(4), 329-333.

Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock, J. R. (2011). The Multi-Attribute Task Battery II (MATBII): Software for human performance and workload research: A user's guide. *NASA Tech Memorandum 217164.*

Sebok, A., Wickens, C., Clegg, B., & Sargent, R. (2014). Using Empirical Research and Computational Modeling to Predict Operator Performance in Novel Situations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58*(1), 844-848

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

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors, In Press.*

Wickens, C. D., Hutchins, S. D., Laux, L., & Sebok, A. (2015). The impact of sleep disruption on complex cognitive tasks a meta-analysis. *Human Factors, In Press*.