Visual attention allocation between robotic arm and environmental process control: validating the STOM task switching model.

Christopher Wickens[1,2], Alex Vieanne[1], Benjamin Clegg[1], Angelia Sebok[2] & Jessica Janes[1]
[1] Colorado State University, Fort Collins, CO [2] Alion Science and Technology, Boulder, CO

Fifty six participants time shared a spacecraft environmental control system task with a realistic space robotic arm control task in either a manual or highly automated version, The former could suffer minor failures, whose diagnosis and repair were supported by a decision aid. At the end of the experiment this decision aid unexpectedly failed. We measured visual attention allocation and switching between the two tasks, in each of the eight conditions formed by manual-automated arm X expected-unexpected failure X monitoring- failure management. We also used our multi-attribute task switching model, based on task attributes of priority interest, difficulty and salience that were self-rated by participants, to predict allocation. An unweighted model based on attributes of difficulty, interest and salience accounted for 96% of the task allocation variance across the 8 different conditions. Task difficulty served as an attractor, with more difficult tasks increasing the tendency to stay on task.

INTRODUCTION

A computational model of multi-task performance in overload called STOM (Strategic task overload management) is designed to predict how task switching will take place, when concurrent performance is impossible and sequential task management is a necessity (Wickens, Gutzwiller & Santamaria, 2015). The model predicts the frequency with which people will decide to switch away from an ongoing task, and then how they choose a task from among a set of alternative tasks that are "waiting in the wings" to be performed. The STOM model was specifically designed to characterize the astronaut in overload circumstances following an unexpected, life-critical failure of an automation device in space. But the model could also apply to any number of circumstances of emergency response or performance when workload is "over the red line". Similar models have been developed by Salvucci & Taatgen (2011) and Freed (2000). However the former also accommodates some parallel processing while the latter has not been formally validated. Our STOM model is closely related to Freed's model.

STOM is a multi-attribute decision model which asserts that that, under high multi-task workload, the tasks that may be switched to, avoided, or those that may be subject to unwarranted "cognitive tunneling" (Dehasise et al, 2011; Wickens & Alexander, 2009; a reluctance to switch away) can be predicted on the basis of each tasks' ranking on each of four critical **task attributes.** In combination, these four attributes can determine the net "attractiveness" (to be switched to or continued), or its inverse "repulsion" (to be avoided, or abandoned rapidly after only a short period of performance). These four attributes are:

Priority: established through mission analysis. For example a safety-critical task, such as maintaining stability in an aircraft (keeping it from stalling) should be of higher priority than one of communicating with air traffic control (Schutte & Trujillo, 1996; Helleberg & Wickens, 2003).

Difficulty: Here, on the one hand, empirical data show (e.g., Arrington & Logan 2004; Kool et al, 2011), and intuition supports the conclusion that easier tasks tend to be more "attractive" than more difficult ones. "I'll get this little task out of the way first, before I tackle the hard job". Such a view is compatible with an inherent "effort-conserving" approach that people may apply in busy circumstances (Kahneman, 2011; Wickens, 2014). On the other hand, we can also identify a counteracting tendency for a more difficult task to be "more attractive" once it has been switched to, and is now an ongoing, rather than alternative task. This tendency was revealed in the meta-analysis of task switching performed by Wickens, Gutzwiller & Santamaria (2015). Such a tendency may be related to a "sunk cost" of staying with a task until it is completed, or the fact that greater difficulty may result from higher working memory demands, which would be sacrificed if it were temporarily abandoned.

Interest, or "engagement". This attribute has been less examined in multi-task workload overload research, but would seem to be operating, for example, in the behavior of a driver who becomes so engaged in an interesting cell phone conversation, that he fails to switch attention to the task of monitoring the roadway for unexpected hazards (and collides with one of them; Horrey & Wickens, 2006). In this example, interest would seem to trump priority, since avoiding a collision is clearly of higher priority than conversing Spink (2006) found the prominent role of interest in task selection, although this study was not carried out in a multi-tasking high workload environment, where STOM is designed to be most applicable.

Salience. This attribute is explicitly defined as the ability of the arrival of a task to "call attention to itself", so, for example, an auditory task (phone rings) would do better (higher salience) than a visual task (message pops up on computer screen) at drawing attention away from an ongoing task. However, both of these sensory attributes are more salient than tasks depending solely on prospective memory (Loukopoulis et al., 2009), such as the pilot needing to remember to lower the landing gear at a specific time.

As noted, each of these four attributes are said to have a "polarity" governing their attractiveness (a high priority, easy, interesting, and salient task will be switched to frequently, and may be slow to leave once it is ongoing). But these attributes may differ in their "weights", and hence how they trade off against one another. For example, will a high priority difficult task "trump" a lower priority easier one?

An important parallel can be drawn between the STOM model of attention (task) switching and the SEEV model of visual attention switching (Wickens, 2015). Both have in common three parameters that determine the attractiveness of a source:

- ***Salience*** of a display or event in SEEV, of the onset of a task in STOM
- ***Value*** of an information source in SEEV; priority of a task in STOM
- ***Effort*** of moving the eyes in SEEV, and difficulty imposed by task performance in STOM.

As an aside, ***Expectancy*** of an event is unique to SEEV, while interest in a task is unique to STOM.

In addition to these attribute parallels, both models are single channel queuing or decision models of where to look or what to do, that do not accommodate concurrent processing. While in many environments there is a dissociation between where one is looking and what task one is doing, it is also true that in many visually distributed work environments the two processes are closely coupled. Our present dual task environment is designed that way with spatially separated tasks; and because of this it is possible to use visual attention as a proxy for task attention, as we do here.

A prior attempt to validate STOM attributes for alternative task attractiveness by Gutzwiller, Wickens and Clegg (2014), required participants to manage sequential performance of the four tasks of the NASA MATB battery: tracking, resource management, visual monitoring and auditory communications. The results indicated that easier tasks were consistently more attractive as alternative tasks, that interest and salience played a secondary role, and that priority appeared to have little influence on task attractiveness.

While the MATB tasks employed in this Gutzwiller et al. (2014) study were more realistic than those examined in many basic switching studies (e.g., Arrington & Logan, 2004), they still remained relatively abstract versions of "real" astronaut tasks. In contrast, the current experiment examines the validity of STOM prediction with two more realistic astronaut tasks, and with considerably greater task training given to well paid, high aptitude volunteers. The two tasks employed were:

- A relatively realistic spacecraft environmental control task, in a simulation called AutoCAMS (Manzey et al., 2012), in which the operator is responsible for managing the mixture of process variables such as oxygen and nitrogen, and fixing minor failures in the system with the assistance of an automated decision aid, called AFIRA. After 6 scenarios in which the decision aid provided correct and useful advice, on the 7th trial it unexpectedly "failed" to appear, and operators unexpectedly (and abruptly) were left to their own knowledge and acquired skills to repair the failure, causing an unexpected abrupt and large workload transition.
- A robotic arm control task, based on the realistic training simulator used at NASA called BORIS, in which our participants manipulated the trajectory of an imaginary astronaut, engaged in repair activity and attached to the end of the arm, along a 3D path (Li et al; 2014; Wickens, Sebok et al; 2015). This task could either be highly automated ("easy" version) or needs to be done by full manual control (hard version). Both modes required some degree of operator intervention, in switching camera views or adjusting the arm movement rate.

The two tasks were required to be performed as well as possible within a restricted time interval of several minutes. However, adhering to NASA safety procedures when operating such an arm (particularly with an astronaut attached), participants were instructed to halt movement on the arm whenever extensive attention (i.e., more than monitoring glances) needed to be devoted to the process control task of AutoCAMS, thereby mandating sequential processing between the two tasks.

In the present paper, our focus is on the pattern of attention switching between them, as predicted by STOM, during the eight conditions defined by crossing normal versus AFIRA-supported operation of AutoCAMS, pre-failure monitoring versus post failure management, with manual versus autopilot operation of BORIS. The focus on task switching, rather than performance was implemented because the environment was one in which the two tasks were primarily performed sequentially in any case (the context in which STOM is relevant). Importantly, our measures of attention allocation between the two tasks were accomplished through two independent techniques: head tracking and control activity.

METHODS

Fifty four participants were recruited and paid $45.00 for their participation in the experiment which ran two sessions and lasted approximately 4 hours. All participants were engineering students or graduate students in psychology.

Robotic Arm Task. The robotic arm task is like that described in Li et al. (2014). In brief, the task required the participant to move, or supervise the movement of, a simulated robotic arm in a series of 3-segment staple patterns. These included a vertical, single axis movement up above a table, a turn above the table top, a horizontally diagonal movement across the top, another turn, and vertical movement to descend to a point at the other side of the table. When the trajectory had been completed, the movement was reversed and the pattern completed in reverse. This cycle continued until the AutoCAMS scenario was completed.

Movement was controlled by two hand controllers. One controller was used to control X, Y, and Z movements in 3D space in which a twist controlled changes in vertical movement, and the other controlled speed (fast or slow) and rotation of the arm. In the manual mode, the ideal staple trajectory was indicated by a 3 dimensional line path to be followed. In the easier automated mode, this same trajectory was executed by an autopilot. In both conditions, the operator was responsible for manually reducing speed as corners were approached, for assuring that the trajectory avoided hazards, and for following guidance to select and change appropriate camera viewpoints of the workspace. Furthermore, in both conditions participants were requested to stop arm movement

when attention for more than a glance, was directed to the AutoCAMS task.

AutoCAMS Task. In this task participants monitored the fluctuating levels of process variables, and were called upon to diagnose and repair occasional failures of the system, such as leaks or stuck valves. Until trial 7, this fault management process was supported by a decision aid called AFIRA (Manzey et al., 2012). However on the 7th scenario, with no warning, AFIRA failed to provide diagnosis and management support and the subjects needed to apply their procedural knowledge, acquired from the previous AFIRA supported diagnoses and repairs and from previous training. They interacted with AutoCAMS using mouse clicks.

The two tasks were configured in the 3-screen layout subtending a visual angle of approximately 120 degrees. The two screens on the right supported the BORIS task, with the rightmost screen providing the 4 camera views necessary to support all arm trajectory motion, while the left BORIS screen (middle of the three screens), provided primarily arm mode control information. The left screen was devoted to AutoCAMS. Operators were requested to sit at a fixed chair location, with their back to the chair and head upright, to maintain this relatively constant visual angle. An Xbox Kinekt head tracker was located above the center screen, to track the allocation of visual attention (assessed here by neck rotation) to the two tasks.

Procedures and Instructions. Participants signed consent forms and were then instructed on how to perform the AutoCAMS task and the BORIS task. Each instructional set involved a series of power point slides, and then several trials, blocked for the two tasks, under close experimenter supervision and guidance to assure that the tasks were performed correctly, and to provide what the experimenter considered to be adequate practice to move into the dual task performance phase. The practice session was 2 hours in duration, with approximately half of the time allocated to each of the two tasks.

Seven dual task trials were then presented on a separate day. Each trial lasted approximately 6 minutes. During each trial, the AutoCAMS system ran normally, until a "routine abnormality" (supported by AFIRA) occurred, sometime between 1 and 3 minutes into the trial. A second phase required the participant to diagnose and repair the failure, lasting approximately 90 seconds (but contingent upon the skill of the participant), and after the repair was completed, the remainder of the trial continued the monitoring requirements. Trial 7 differed from trial 6 in that the AFIRA aid was, unexpectedly not present, an event which participants were never instructed could occur. The AFIRA box was present to indicate the presence of a failure, but no diagnosis or management advice was available. It was assumed that to the extent that participants had become reliant upon AFIRA during training and the first 6 trials to assist diagnosis and system repair, they would find themselves in an unexpectedly high workload period in the failure phase of trial 7 (Wickens, Vieane, Clegg & Sebok, 2015). Trial 7 lasted 10 minutes and the failure was introduced 3 minutes into the trial.

In all trials, participants continued the staple cycle on BORIS, subject to the sequential task constraints, until the AutoCAMS trial was complete. They could voluntarily switch between tasks whenever they chose.

The participants were randomly assigned to either a manual or an autocontrol BORIS condition. Independent of their assignment, all participants were explicitly and clearly instructed that both tasks were equally important. Thus participants in the autopilot BORIS condition were clearly reminded of the criticality of speed control, hazard monitoring and camera selection, even though their attention was not required for actual arm control. Furthermore, these participants had been trained in some aspects of manual control, in case that should be required.

After the final scenario, participants provided the four attribute ratings of the two tasks along a 5 point scale.

RESULTS

Attention Measure

Figure 1 presents the percent time the eye spent looking at the AutoCAMS display, as a function of AutoCAMS phase (monitoring versus failure management). This is the perfect inverse of the time spent looking at the two BORIS displays. The upper line are data from the BORIS autopilot subjects, and the lower line are data from the BORIS manual subjects. The left graph shows data from trial 6 (routine abnormality management) and the right graph is from trial 7 (unexpected AFIRA failure).

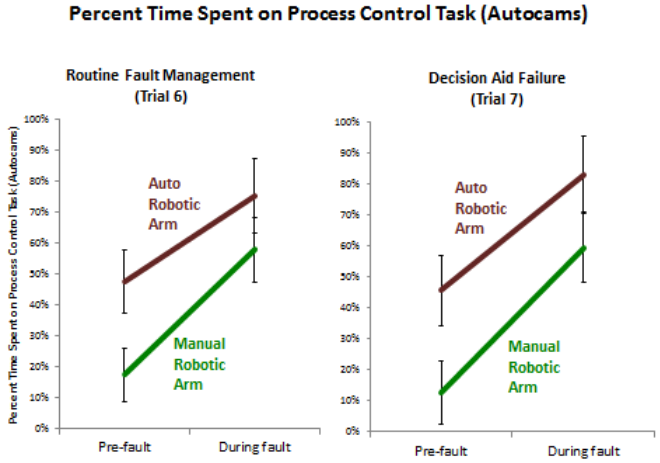

Figure 1: Percent time of visual fixation on the AutoCAMS display.

Separate mixed model ANOVAs were carried out on each trial (graph), because some data recording failures left fewer data points available for trial 7. The graphs clearly indicate common trends on both trials. There was a significant increase in attention to AutoCAMS during failure management compared to pre-failure monitoring ($F=73.9$, trial 6; $F=58.2$, trial 7; both $p<.01$). There was also a significant increase in attention to AutoCAMS when the concurrent BORIS task was in its easier autopilot mode, relative to its more demanding manual model (Trial 6, $F=15.5$; Trial 7, $F=19.0$; both p's$<.01$). The two variables did not interact on either trial.

While the general pattern is similar on both trials 6 an 7, one difference is statistically and practically significant: for the autopilot group during fault management, a pairwise comparison revealed that attention to AutoCAMS was significantly greater on trial 7 (83%) compared to trial 6 (73%). ($t=3.93$ $p<.01$), reflecting the significantly greater demands of the unexpected loss of decision support.

While not reported here, we observed that the measure of mouse clicks on AutoCAMS, reflecting task (rather than visual) attention very closely mirrored that of visual attention, with the same significant effects. The two attention measures were closely correlated ($r=0.845$) across the 8 data points.

MODELLING

Participant ratings of the two different tasks across the four different STOM attributes in the eight different conditions of figure 1 are shown in the three middle columns of Table 1 for AutoCAMS (left number) and BORIS (right number). Priority, while rated, was not employed in the model and is not depicted here. The right column depicts the total attractiveness predicted by STOM as discussed below.

Table 1: Attribute ratings. The first value in each column is the rating for AutoCAMS. The second is the rating for BORIS. Lower ratings indicate greater attractiveness.

Condition	Interest	Salience	Difficulty	I+S+D
MANmon6	2.9 1.6	3.3 3.2	3.6 2.5	*9.8* 7.3
MANmon7	2.9 1.6	3.3 3.2	3.6 2.5	*9.8* 7.3
MAN fail 6	2 1.6	2.4 3.2	2.8 2.5	*7.2* 7.3
MAN fail 7	2 1.6	2.4 3.2	1.9 2.5	*6.3* 7.3
ATP mon6	2.9 2.8	2.4 3.3	3.6 4.0	*9.8* 10.0
ATP mon7	2.9 2.8	3.3 3.2	3.6 4.0	*9.8* 10.0
ATP fail 6	2 2.8	2.4 3.2	2.8 4.0	*7.2* 10.0
ATP fail 7	2 2.8	2.4 3.2	1.9 4.0	*6.3* 10.0

In considering the ratings, three important caveats should be noted. First, only difficulty was asked to be rated differently between the two AutoCAMS phases. Second, participants provided only one rating at the end of the third trial block, and this block contained both trial 6 (AFIRA on) and trial 7 (AFIRA gone). However we were able to infer that task difficulty was approximately 50% higher (ratings, 67% lower) in this failure stage on trial 7 versus trial 6, on the basis of analysis of workload carried out on the corresponding trials in Wickens, Clegg, Vieane & Sebok (2015).

Finally, the differential attribute ratings of AutoCAMS between the pre-disturbance (monitoring) phase and the during disturbance (management phase), shown in Table 1 were based only on the ratings of 11 of our subjects, as the remaining subjects only provided a single rating of Interest and Salience for the full AutoCAMS trial. We assumed that the ratings provided by these 11 were a random sample, and hence typical of the remaining subjects, an assumption validated by the high correlation ($r=0.94$) of those attributes that were rated by both of the two groups. The correlation between the group of 11 and the full cohort in attention allocation % across the eight conditions was also 0.94, indicating that their attention distribution pattern was consistent with that of the larger population.

The attribute rating data indicated that requiring manual control in BORIS rendered it to be judged significantly more interesting (1.6 vs. 2.8; $t = 3.39$. $p<.01$), more difficult (2.5 vs. 4.0; $t = 5.27$, $p<.01$), and more attractive overall (7.3 vs. 10.0). AutoCAMS was rated more difficult during fault management than pre failure monitoring, and in the fault management condition (applying the heuristic from the Wickens, Clegg, Vieane & Sebok (2015) study), AutoCAMS was rated more difficult on trial 7 (AFIRA gone) than trial 6.

The original version of the model predicted the attractiveness of an alternative task to be a linear equally weighted combination of its attributes: Attractiveness = P + I + S – D, such that the higher value would make the task more likely to be switched to. However in our modeling exercise (See Sebok et al. 2015 for details) two facts became evident: First, priority played little role (replicating the negative findings of Gutzwiller et al., 2014). Second, because only two tasks were performed, there was never any choice *between* alternative tasks, a condition necessary for the "easy task preference" to play a role as suggested by the original STOM model and meta-analysis (Wickens, Gutzwiller & Santamaria, 2015). Thus the role of task difficulty differences (e.g., between the easier autopilot and harder manual BORIS control) became ambiguous. Would the harder task be stayed on longer once it was chosen as revealed by the original meta-analysis? Or would it become less attractive when it was an alternative task? Here we let the data in figure 1 speak for themselves. In every case, tasks of greater difficulty (manual BORIS, failure management, and unexpected AFIRA withdrawal) had longer periods of attention. Thus we let difficulty be an attractor rather than a repelling factor, and the model became:

$$\text{Attractiveness} = I + S + D.$$

From the net attractiveness values of each task, (right column of Table 1), we computed a difference which was the extent to which AutoCAMS was favored over BORIS.

Figure 2 depicts the scatter plot of the attractiveness of AutoCAMS that gives the predicted percent allocation, against the proportion of empirically observed visual attention to AutoCAMS (the eight data points in Figure 1) and indicates a strong degree of model prediction fit across the eight conditions with $r = 0.979$, or accounting for 96% of the variance. A correlation with AutoCAMS click activity was 0.92, indicating that task attention was also well predicted.

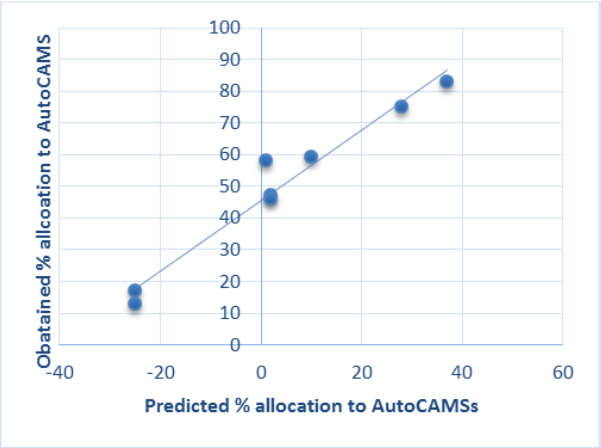

Figure 2: Scatter plot of predicted vs. actual AutoCAMS attractiveness

DISCUSSION

In this study two realistic astronaut tasks were time shared to provide data to validate STOM, the multi attribute decision model of sequential task switching and management. In order to create data points predicted to be of varying attention allocation, the robotic arm task was varied in difficulty through the degree of automation, and the environmental control task was varied both by phase (monitoring versus failure management) and, within the latter, by the support or unexpected withdrawal of a diagnostic decision aid. These difficulty manipulations also influenced salience and interest.

Together we found that an equal weighting attribute model of I + S + D accounted for over 95% of the variance in visual attention allocation across the eight conditions. Two important changes to the original STOM model of Wickens, Gutzwiller & Santamaria (2015) were implemented. First, priority appears to play a minimal role, and its inclusion in some versions of the model actually degraded model fit. This result is somewhat puzzling, since priority does play a role in concurrent multi-tasking. But, in contrast to salience, difficulty, and interest, which are constantly at the forefront of the subject's experience while performing a task, priority remains a step removed. Priority is identified through pre-task instructions, but it is neither explicitly nor implicitly reminded during the course of the experiment. Second, the role of difficulty seems to possess a kind of hysteresis: when tasks are waiting to be performed, we select the easier. But once selected, more difficult tasks tend to be "stickier" and more resistant to switching away from. Hence we might expect more cognitive tunneling on more difficult tasks.

Our modeling focused on the mean attribute ratings and attention allocation over participants. In subsequent analyses we examined whether parameterizing data to individual participants might provide better fits (e.g., because different participants find the tasks of different interest). Here we found that the mean correlation for individual fits is 0.96, roughly equivalent to the 0.979 correlation of the means.

ACKNOWLEDGEMENTS

This work was supported by NASA Grant NNX12AE69G. Dr. Brian Gore was the scientific/technical monitor. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of NASA. Nathan Herdener, a PhD student at CSU, provided essential methodological support.

REFERENCES

Arrington, C. M. and Logan, G. D. (2004). The cost of a voluntary task switch. *Psychological Science, 15*, 610-615

Dehais, F., Causse, M., & Tremblay, S. (2011). Mitigation of conflicts with automation. *Human Factors, 53,* 448–460.

Freed, M. (2000). Reactive Prioritization. *Proceedings 2nd NASA workshop on planning & scheduling in space*. Wash D.C.: NASA

Gutzwiller, R., Wickens, C. & Clegg, B. (2014) Workload overload modeling, In *Proceedings of the HFES Annual Meeting*.

Horrey, W. & Wickens, C.D. (2006), the impact of cell phone conversations on driving. *Human Factors*, 48, 196-208.

Jin, J., & Dabbish, L. 2009. Self-interruption on the computer: a typology of discretionary task interleaving. In: *Proceedings of the 27th Conference on Human Factors in Computing systems*, *CHI '09*, 1799–1808.

Kahneman, D. (2011) *Thinking Fast and Slow*. NY: Farrar Strauss & Girioux.

Kool, W., McGuire, J. T., Rosen, Z. B., and Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, 139, 665-682

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: *Journal of Cognitive Engineering and Decision Making*, 6, 1–31.

Li, H. Wickens, C., Sarter, N., & Sebok, A. (2014) Stages and Levels of Automation in Support of Space Teleoperations. *Human Factors.*

Loukopolis, L., Dismukes, K. & Barshi, I. (2009), *The Multi-tasking Myth*. Averbury Vt. : Ashgate..

Salvucci, D. & Taatgen, N (2011). The Multi-tasking mind. Oxford U. Press.

Schutte, P. C., & Trujillo, A. C. (1996). Flight crew task management in non-normal situations. *Proceedings of the Human Factors 40th Annual Meeting,* 244-248.

Sebok, A., Wickens, C., Sargent, R., Clegg, B., & Jones, T. (2015) *Space Performance Research Integration Tool (S-PRINT): Development and Validation of a Model-Based Tool to Predict, Evaluate and Mitigate Excessive Workload Effects - Year 3 Status Report / Final Report*. Delivered to the NASA Human Research Program under Grant NNX12AE69G.

Spink, A., Park, M., and Koshman, S. (2006). Factors affecting assigned information problem ordering during Web search: An exploratory study. *Information Processing & Management, 42*, 1366–1378.

Payne, S. J., Duggan, G. B., and Neth, H. (2007). Discretionary task interleaving: Heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology: General*, 136, 370-388.

Wickens, C.D (2014) Effort in human factors performance and decision making. *Human Factors, 56,* 1329-1336

Wickens, C.D. (2015). Noticing events in the visual workplace: The SEEV and NSEEV models. In R. Hoffman & R. Parasuraman (Eds). *Handbook of Applied Perception*. Cambridge, U.K.: Cambridge University Press

Wickens, C., Clegg, B. Vieanne, A. & Sebok, A. (2015) Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors*.

Wickens, C. Gutzwiller, R., & Santamaria, A. (2015,) Discrete task switching in overload. *Intl Journal of Human Computer Studies. http://dx.doi.org/10.1016/j.ijhcs.2015.01.002*

Wickens, C. Sebok, A., Li, H., Sarter, N. & Gacy, A. (2015) Using Modeling and Simulation to predict Operator Performance and Automation-induced Complacency with Robotic Arm Automation. *Human Factors*.