Dual-Task Performance Consequences of Imperfect Alerting Associated With a Cockpit Display of Traffic Information

Christopher Wickens and **Angela Colcombe,** University of Illinois, Champaign, Illinois

Objective: Performance consequences related to integrating an imperfect alert within a complex task domain were examined in two experiments. **Background:** Cockpit displays of traffic information (CDTIs) are being designed for use in airplane cockpits as responsibility for safe separation becomes shared between pilots and controllers. Of interest in this work is how characteristics of the alarm system such as threshold, modality, and number of alert levels impact concurrent task (flight control) performance and response to potential conflicts. **Methods:** Student pilots performed a tracking task analogous to flight control while simultaneously monitoring for air traffic conflicts with the aid of a CDTI alert as the threshold, modality, and level of alert was varied. **Results:** As the alerting system became more prone to false alerts, pilot compliance decreased and concurrent performance improved. There was some evidence of auditory preemption with auditory alerts as the false alarm rate increased. Finally, there was no benefit to a three-level system over a two-level system. **Conclusion:** There is justification for increased false alarm rates, as miss-prone systems appear to be costly. The 4:1 false alarm to miss ratio employed here improved accuracy and concurrent task performance. More research needs to address the potential benefits of likelihood alerting. **Application:** The issues addressed in this research can be applied to any imperfect alerting system such as in aviation, driving, or air traffic control. It is crucial to understand the performance consequences of new technology and the efficacy of potential mitigating design features within the specific context desired.

INTRODUCTION

In a variety of human-system integration contexts, humans are asked to act in parallel with automation as diagnostic systems, discriminating events from nonevents (Dixon & Wickens, 2006; Getty, Swets, Pickett, & Gonthier, 1995; Madhavan, Wiegmann, & Lacson, 2006; Maltz & Shinar, 2003; Meyer, 2004; Metzger & Parasuraman, 2005; Sorkin & Woods, 1985; Wickens & Dixon, 2007). In the current context, our interest is in parallel human and automation performance in airborne conflict detection/alerting systems (Xu, Wickens, & Rantanen, 2007).

Such alerting systems have several important and complex properties that are affected by their design and, hence, affect the joint performance of the human-automation "team" as these were first studied by Pollack and Madans (1964) and then analyzed in detail by Sorkin and Woods (1985). Within the context of signal detection theory, two important influences or impacts may be considered: the reliability, or *sensitivity,* of the automated system in discriminating events from nonevents, and the threshold setting, or *response bias,* of the alerting system, dictating the ratio of the two kinds of automation responses (silent, alarm) and, therefore, the two kinds of automation errors (misses vs. false alerts).

Regarding sensitivity (determined by the reliability and quality of the algorithms that process the raw data in the external world), one would imagine that improved automation sensitivity would improve sensitivity of the human-automation

Address correspondence to Christopher D. Wickens, Micro Analysis and Design, 4949 Pearl East Circle, Suite 300, Boulder, CO 80301; cwickens@alionscience.com. *HUMAN FACTORS,* Vol. 49, No. 5, October 2007, pp. 839–850. DOI: 10.1518/001872007X230217.

system as a whole, a phenomenon that has been well documented (e.g., Dixon & Wickens, 2006; Maltz & Shinar, 2003).

More complex is the relationship between reliability and human *dependence* on the automation system (e.g., agreement with automation advice; Wiegmann, 2002). Results indicate that with high-reliability automation (reliability that is greater than that of the human alone), total system performance is improved above the capabilities of the human alone (but less than total dependence on the automated system would dictate; Parasuraman, 1987). Then, as reliability degrades, humans also become less dependent, but even as reliability drops below a threshold at around $r = .75$ ($d' = 1.35$), humans may continue to depend on the imperfect diagnostic automation, even if their performance would be better if this automated advice were ignored (Maltz & Shinar, 2003; Wickens & Dixon, 2007).

Further complicating the picture are two additional factors: (a) Dependence may be altered as a function of how salient or obvious the automation errors are (Wiegmann, 2002; Madhavan et al., 2006). (b) Of interest in the current study is that there is a differential impact of automation misses and automation false alarms. That is, humans may trust or depend on the automation more to the extent that it detects events (leading to a higher hit rate, but also a higher false alarm rate) or to the extent that it is more often silent (leading to a greater miss rate).

Meyer (2001, 2004) analyzed the cognitive processes related to these two threshold settings and defined *compliance* as the state when the alarm sounds and *reliance* as the cognitive state when the alarm is silent (signaling "all is well"). The distinction is important because the ratio of automation detection to automation silent responses (influencing the respective frequency of false alarms vs. misses when automation is imperfect) will lead to

- decreasing automation dependence when the automation sounds (decreasing compliance with the alert – the "cry-wolf" phenomenon; Getty et al., 1995) as automation false alarm rate increases; and
- decreasing automation dependence when the automation is silent (decreasing reliance upon the alerting system) as automation miss rate increases.

These findings tend to be supported by the data (Dixon & Wickens, 2006; Maltz & Shinar, 2003).

Although the false alarm and miss rates of automation may be manipulated independently (Dixon & Wickens, 2006; Maltz & Shinar, 2003), in practice they are more often varied in a negatively correlated fashion, as the threshold of the automated system is varied (Getty et al., 1995; Levinthal & Wickens, 2005; Swets, 1991; Wickens, Dixon, Goh, & Hammer, 2005). Indeed, designers typically place the threshold low (low beta) in order to guard against automation misses, which are typically assumed to be more costly than automation false alarms (Getty et al., 1995).

However, a point not always realized by designers is that high false alarm rates, creating the cry-wolf effect (Breznitz, 1983), can lead the human operator to ignore automated advice, thereby compromising the effectiveness of the combined human-automation system (Maltz & Shinar, 2003). Indeed, Sorkin and Woods (1985) have shown that the combined effectiveness of human and automation may be greater than individual human effectiveness only when this false alarm rate is relatively low (the threshold is high). Adding to the complexity of the picture is the fact that a low base rate of hazard events to be detected, typical in many real-world environments, will further escalate the false alarm rate if the miss rate is also to be kept low (Getty et al., 1995; Krois, 1999; Parasuraman, Hancock, & Olofinboba, 1997). Indeed, in some air traffic control conflict detection systems, the probability that an alarm will be false may be well over 0.50 (Krois, 1999).

As the alert threshold is varied, then, the states of reliance and compliance will covary negatively, with the reliance being high and the compliance low, as the alert threshold is set to the typical low value (e.g., with high false alarm rates). Importantly, such variation has implications for attention allocation and concurrent task performance in the multitask domains in which alarms have proven to be most effective (Getty et al., 1995; Wickens & Dixon, 2007). The high reliance induced by a low miss rate should assure operators that the system will alert them if a true failure occurs, and hence they will allocate plenty of residual attention to concurrent tasks (Wickens, Dixon, Goh, et al., 2005). The high false alarm rate could lead to operators ignoring many of those false alarms entirely and, hence, continuing to support concurrent tasks. However, more likely, those alarms (both true and false) will be checked eventually, although after a cry-wolf delay. Hence the increased number of alarms with the lower threshold will increase the total number of interruptions and might harm the concurrent task as much as, or even more than, the miss-prone system would.

Indeed, these two offsetting trends of lowering the threshold, leading to (a) less concurrent task disruption (because of more reliance when the alarm is silent) and (b) more disruption (because of an increasing number of alarms), seem to be reflected by the relatively ambiguous pattern of concurrent task performance that has been observed in dual-task experiments as alert threshold has been varied (Dixon & Wickens, 2006; Dixon, Wickens, & McCarley, 2007; Levinthal & Wickens, 2005; Wickens, Dixon, & Johnson, 2006; Wickens, Dixon, Goh, et al., 2005; see Wickens, Dixon, & Ambinder, 2006, for a summary). A goal of the current study is to examine the effect of varying the alert threshold of an airborne conflict alerting system on processing the alert itself and on concurrent task performance.

In addition to its threshold level, a second characteristic of the alert system that may influence how attentional resources are distributed between the alerted task and the ongoing task or tasks is the modality of the discrete alert. Auditory presentation has typically been the modality of choice for such alerts, but its attention-capturing properties (Spence, 2001) can be disruptive to concurrent ongoing tasks, leading to rapid processing of the auditorily alerted task (compared with a visual alert) but greater disruption of ongoing interrupted tasks. Such a disruption could be serious if the ongoing task is of high priority (e.g., flight control).

Whereas auditory preemption theory predicts this asymmetric effect of modality on ongoing task versus alerted task (Iani & Wickens, 2007; Wickens & Liu, 1988), multiple-resource theory (Wickens, 2002; Wickens & Hollands, 2000) predicts a symmetric benefit, such that both tasks will benefit from an auditory (relative to a visual) alert presentation when the ongoing task is itself visual. The differing predictions of these two theories will be evaluated.

A third factor of the alert that was manipulated in our experiments is the nature of the alarm itself. Two-state "on-off" alarms are frequently used in alerting situations, but cogent arguments have been offered for the benefits of three (or more) state *likelihood alarms* that can self-report their own level of confidence that a dangerous state exists (Latorella, 1996; Sorkin, Kantowitz, & Kantowitz, 1988; Sorkin & Woods, 1985; St. John & Manes, 2002; Woods, 1995). Although Sorkin et al. (1988) found no impact of likelihood alerting on concurrent tracking performance, they did find that when the ongoing (tracking) task difficulty was high, a likelihood alert supported better performance on the alerted task. St. John and Manes (2002) found that visual likelihood alert information improved accuracy on a search task, but they did not impose a concurrent task. Together these studies suggest that likelihood alerting will likely have some positive impact on alerted task performance.

Finally, the fourth factor we varied was the difficulty (stability) of the concurrent tracking task, in order to establish the robustness of the observed effects of the other three variables across levels of workload.

The context for our evaluation of these four factors is the cockpit display of traffic information (CDTI), a system proposed within future cockpits to provide pilots with a traffic display that is partially redundant with the air traffic controller's display and may, in some future airspace plans, allow pilots to monitor their course for conflicts and to initiate the choice to make route changes (Johnson, Battiste, & Bochow, 1999; Thomas & Wickens, 2005; Wickens, Helleberg, & Xu, 2002; Wickens, Goh, Helleberg, Horrey, & Talleur, 2003). Indeed, it has been proposed that such systems be coupled with discrete alerts (Thomas & Rantanen, 2006; Xu et al., 2007), paralleling similar alerting systems for air traffic control (Metzger & Parasuraman, 2005) and for more emergency airborne conflicts (the Traffic Alert and Collision Avoidance System [TCAS]).

A challenge for the CDTI alerting system is that it is more strategic in nature than the tactical TCAS, thereby imposing a longer look-ahead time. But this longer time will induce greater uncertainty as to the future state of potentially intersecting trajectories (Kuchar & Yang, 2000), hence considerably lowering the reliability of the alert system (its "sensitivity" in signal detection terms), and hence will amplify the potential false alarm (FA) problems described previously. A related feature of such detection systems is that the longer a discrete detection response is delayed (longer response time [RT]), the more reliable the information will become, and hence the more accurate the response is likely to be.

Thus the current study was intended to inform both the specific avionics design community, as well as the more general audience of alert researchers, regarding the joint effects of the four independent variables: the modality of the alert; the nature of the alert (likelihood vs. binary); the difficulty of an ongoing task; and (between experiments) the alert threshold, from a neutral setting (equal frequency of automation misses and FAs) to an FA-prone setting, more typical of operational alerting systems.

We hypothesized the following:

1. The FA-prone automation will reduce compliance with the automated system and, therefore, increase the time (RT) for pilots to switch attention to the alerting system and detect conflicts (cry-wolf effect). However, a delayed response may actually improve conflict detection accuracy.
2. In contrast, the FA-prone system will increase reliance (because of a reduced number of automation misses) and therefore should improve concurrent task performance because visual resources would not need to be allocated to monitoring the raw traffic data when the alert is "silent." This effect may not be strong, however, as it could be offset by the greater interruption frequency of the FA-prone system.
3. Auditory alerts should improve performance on the alerting task, but should (a) degrade the concurrent task to the extent that preemption theory is operating and driving attention rapidly to the automated domain or (b) improve the concurrent task to the extent that multiple resource theory is operating and thus allow the pilot to capitalize on two (visual and auditory) perceptual resources.
4. Performance on the concurrent task should improve with the three-state likelihood alert, relative to the two-state alert, because the likelihood alert, by providing more specific information about the severity of a given threat, should allow pilots to more optimally distribute attention between the two tasks. The likelihood alert may lead to slower responses to alerts (especially midlevel, less serious alerts) in order to preserve ongoing task performance. However, conflict detection accuracy should not suffer.

EXPERIMENT 1

Methods

Twelve student pilots from the University of Illinois Institute of Aviation were recruited to participate in the 3-hr experiment. Pilots were paid $9/hr. Figure 1 shows a display of the experimental task. Each pilot completed two sessions of a computer task wherein they performed an ongoing, first order, compensatory tracking task with a bandwidth of 0.30 Hz that was presented centrally on the computer screen. Simultaneously, pilots monitored for air traffic conflicts on a CDTI with the aid of an imperfect automated alert. The tracking task required pilots to keep a cursor within an acceptable position inside a target rectangle and was controlled with a joystick using the left hand.

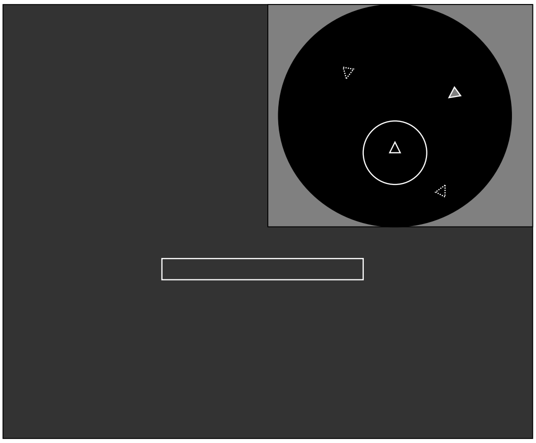

Figure 1. Display for the experimental task in Experiments 1 and 2 with the centrally presented tracking task and the CDTI in the upper right corner.

As shown in Figure 1, a simple CDTI display was presented in the upper right corner of the screen. The visual angle between the centered tracking error cursor and the near corner of the CDTI was 6°. The angle to the far corner was 11°. The CDTI monitored for potential collision threats of slowly moving aircraft (all at common altitude) and warned pilots with a visual or an auditory alert if a collision threat (proximity < 3 miles) was predicted. The visual alert was of sufficient salience, illuminating the entire periphery of the CDTI, as to be easily seen when visual attention was focused on the tracking display; thus any costs to visual presentation could not be attributable to scanning differences.

Automated conflict detection was less than perfect (75% reliable, automation $d' = 1.33$), but the ratio of automation FAs to misses was 1:1 (10 misses and 10 FAs, out of 46 conflict and 34 nonconflict, events respectively; see Table 1a). In the likelihood alert condition, the distribution is shown in Table 1b. In Table 1b, the midlevel alert was signaled only when separation was either 3 (conflict) or 4 (nonconflict) miles. Note that each of these is a "difficult to judge" situation, either

TABLE 1: Event Rates for the Two Experiments and Alert Conditions

Automation Response	Conflict	Nonconflict	*p*(H)	*p*(FA)
			Experiment 1: High (Neutral) Threshold (a) Binary	
Yes	36	10	.78	.29
No	10	24		
			(b) Likelihood	
Yes	30	7	Automation Yes = Hit: 30/47 = .64 (high-level)	Automation Yes = Hit: 7/36 = .19
Maybe	10	11	Automation Yes + Maybe = Hit: 40/47 = .85 (high + mid)	Automation Yes + Maybe = Hit: 18/36 = .50
No	7	18	Automation Yes + ½ Maybe = Hit: 35/47 = .74	Automation Yes + ½ Maybe = Hit: 12/36 = .33
			Experiment 2: Low (FA-Prone) Threshold (c) Binary	
Yes	42	16	.91	.47
No	4	18		
			(d) Likelihood	
Yes	30	12	Automation Yes = Hit: 30/43 = .70	Automation Yes = Hit: 12/40 = .33
Maybe	10	10	Automation Yes + Maybe = Hit: 40/43 = .93	Automation Yes + Maybe = Hit: 22/40 = .55
No	3	18	Automation Yes + ½ Maybe = Hit: 35/43 = .81	Automation Yes + ½ Maybe = Hit: 17/40 = .43

Note. For the likelihood condition, we present three different means of calculating hit (H) and false alarm (FA) rate. In the first, the middle ("maybe") category is assigned to an automation "no" response; in the second, it is assigned to a "yes" response; and in the third, half of the "maybe" events are assigned to "yes" and half to "no."

a near miss or a near hit. The d' calculation based on dividing these midcategory "maybe" responses between "yes" and "no" yields a d' of 1.15 (see Table 1b). When pilots detected a conflict, they clicked on the conflict aircraft's icon with the left or right mouse button to indicate the direction the aircraft should be routed to in order to avoid a conflict with their own aircraft.

Pilots were instructed to place slightly greater emphasis on tracking than on the detection task, given the standard hierarchy in which aviating has a higher priority than navigating (Schutte & Trujillo, 1997). In addition, pilots were told that the automation was not perfect and might miss some conflicts and falsely identify nonconflicts as conflicts.

Tracking difficulty (stable vs. unstable), alert modality (visual vs. auditory), and alert type (binary vs. likelihood) were manipulated within subjects. For the visual binary alert, the border was changed to red for conflict trials. For the visual likelihood alert, the border was also changed to amber for the "maybe" near-conflict trials. For the auditory binary alert, the synthesized voice said "conflict conflict." For the auditory likelihood alert, the voice also said "traffic traffic" for the "maybe" trials. Each group of pilots participated in four separate conditions per session, and each condition lasted approximately 14 min. During each condition, a new aircraft appeared on the screen every 10 s in a continuous stream, for a total of 80 traffic aircraft per condition. There were never more than four aircraft icons on the screen at one time.

Conflict generation consisted of a random assortment of conflict angles between 30° and 300°, from the left, from the right, and passing in front of and behind ownship. Of the events, approximately 50% were conflict events, which were manipulated to represent a range of threat seriousness according to closest point of approach. In the likelihood alert conditions, the passage of each intruder was subdivided into three groups: (a) less than 3 miles (conflict); (b) 3 miles (conflict) or 4 miles (nonconflict) from ownship (both of these difficult-to-discriminate scenarios received the midlevel auditory ["traffic"] or visual [amber] signal); and (c) greater than 4 miles (nonconflict). Pilots were to judge a conflict to be any aircraft that would penetrate their protected airspace (3 nautical miles). As shown in Figure 1, a standard 3-mile ring was placed around ownship to support this judgment. Automation reliability was 75%, with errors equally and randomly distributed between automation FAs and misses.

Pilots filled out an informed consent form and instructions and then received three 3-min blocks of practice: one block of tracking only, one block of CDTI monitoring, and one block with both tasks. They then experienced two sessions on consecutive days, one with the likelihood and one with the binary alert. This order was counterbalanced. Within each session, they experienced the four conditions determined by modality and tracking difficulty, again in a counterbalanced order.

Results: Experiment 1

Four dependent variables were analyzed: Time to respond to the conflict (RT), accuracy in discriminating conflicts from nonconflicts (d', measure of sensitivity), tracking error on the concurrent task, and the percentage of dwell time on the tracking task as compared with the CDTI display. Data were tested for skew, and all outliers (less than 1% of the data) were eliminated from the analysis. Repeated measures ANOVAs were executed with alarm type, alarm modality, and tracking difficulty as within-subjects independent variables.

Response time. Pilots had slower response times to CDTI conflicts during unstable tracking as compared with stable tracking, $F(1, 11) = 7.02$, $p < .05$. However, pilots were equally fast to respond to CDTI alerts regardless of alert modality (auditory vs. visual), $F(1, 11) = 0.45$, $p > .10$, or alert type (likelihood vs. binary), $F(1, 11) = 0.47$, $p > .10$. As shown in Figure 2, an interaction emerged between alarm type and tracking difficulty, $F(1,$

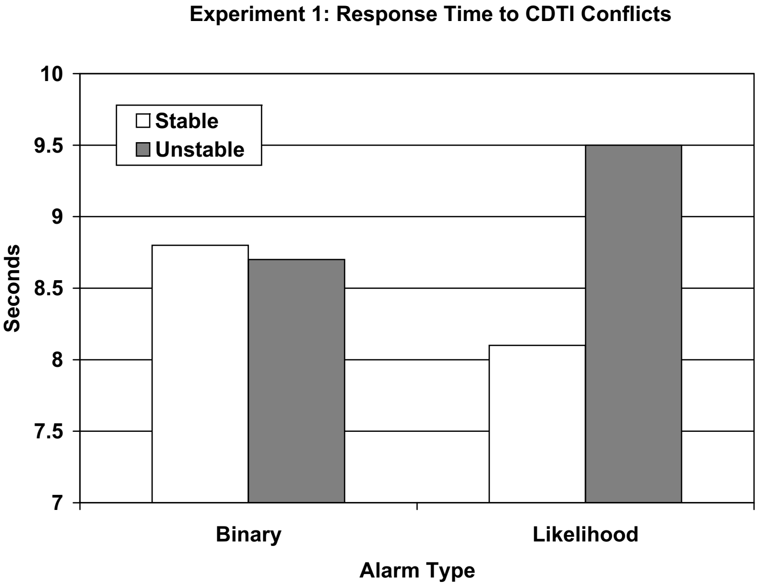

Figure 2. A significant interaction in Experiment 1 between alarm type (binary vs. likelihood) and tracking stability (stable vs. unstable) for response times (in seconds), indicating a benefit for likelihood alerting with stable tracking but a cost for likelihood alerting during unstable tracking.

11) = 8.62, $p < .05$, which indicated a beneficial effect of the likelihood alert with stable tracking but a cost associated with the likelihood alert during unstable tracking. This may be evidence that the likelihood alert allowed for a better distribution of attention.

Sensitivity. Increased tracking difficulty had no impact on pilots' ability to detect CDTI conflicts, as reflected by sensitivity, $F(1, 11) = 2.20$, $p > .10$. Pilots were more accurate in detecting CDTI conflicts in the auditory condition than in the visual alert condition ($M = 2.14$ vs. 1.37), $F(1, 11) = 54.07$, $p < .01$. Alarm type (binary vs. likelihood) did not impact pilots' conflict detection accuracy, $F(1, 11) = 0.14$, $p > .10$. There were no other significant main effects or interactions.

Tracking error. Pilots' tracking error was almost double for the unstable tracking task as compared with the stable tracking task ($M = 193.47$ vs. 353.93), $F(1, 11) = 372.78$, $p < 0.01$. Tracking performance was the same regardless of the modality of the alert, which, coupled with the improvement of conflict detection accuracy with auditory alerts, supports multiple resource theory. Tracking error was worse during the likelihood alerting condition as compared with the binary alerting condition (M = 282.25 vs. 264.5), F(1, 11) = 8.26, $p < .05$. As with the RT variable, there was no interaction between alert modality and tracking difficulty, again supporting multiple resource theory.

As shown in Figure 3, an interaction between alert type and modality emerged, $F(1, 11) = 6.13$, $p < .05$, such that concurrent tracking was particularly hurt by visual likelihood alerts.

EXPERIMENT 2

Method

The method for Experiment 2 was identical to that of Experiment 1 except that the ratio of automation FAs to misses was 4:1 (16 FAs out of 40 nonconflict trials, and 4 misses out of 40 conflict trials) instead of 1:1. Twelve new student pilots were used in Experiment 2 and were statistically the same in experience and demographic variables as those who participated in Experiment 1. The distribution of events is shown in Table 1c (binary) and Table 1d (likelihood). Thus the alert threshold was reduced in Experiment 2 in order to examine the impact of increased FAs (interruptions) and decreased misses on both alerted and concurrent task performance.

Results

Response time. Increased tracking difficulty did not directly affect the time it took pilots to

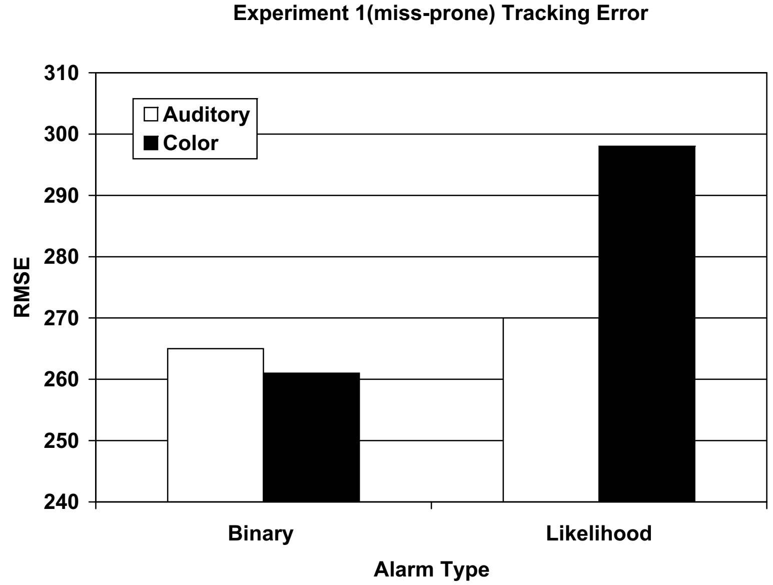

Figure 3. A significant interaction in Experiment 1 for tracking error (RMSE = root mean square error) between alert type (binary vs. likelihood) and alarm modality (auditory vs. visual) indicating a likelihood alarm cost only for visual alerts.

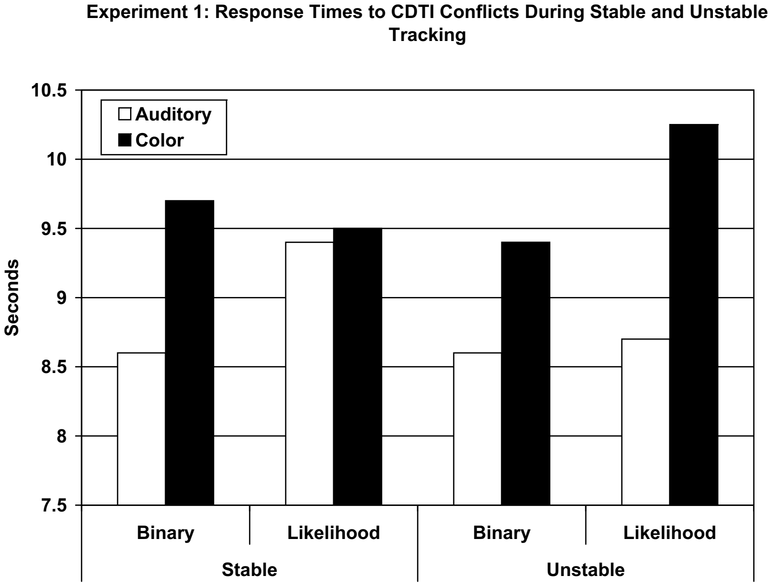

Figure 4. A significant three-way interaction for response time in Experiment 2, wherein a cost emerged for visual likelihood alerts when tracking was difficult and unstable (right) that wasn't present with the easier, stable tracking task.

respond to CDTI conflicts, $F(1, 11) = 2.3, p > .10$. There was a marginally significant effect of alarm modality, with faster responses to auditory alerts ($M = 8.9$ s) than to visual alerts ($M = 9.5$ s), $F(1, 11) = 4.09, p = .07$. There was no main effect of alarm type (binary vs. likelihood) on response times to the conflicts, $F(1, 11) = 0.16, p > .10$. There were no significant two-way interactions among alarm type, alert modality, or difficulty when all responses were grouped. However, as shown in Figure 4, there was a significant three-way interaction among alarm type, modality, and tracking difficulty, $F(1, 11) = 5.2, p < .05$. Only when tracking was difficult (unstable; right side of Figure 4) did a cost for the visual likelihood alarm emerge that wasn't present with the stable tracking task.

Sensitivity. Increased tracking difficulty had no direct impact on participants' sensitivity, $F(1, 11) = 0.24, p > .10$. However, tracking difficulty did interact with alarm type, $F(1, 11) = 4.69, p = .05$, such that there was a likelihood alarm cost to sensitivity with difficult tracking. Pilots were marginally more sensitive to auditory alerts ($M = 3.41$) compared to visual alerts ($M = 3.34$), $F(1, 11) = 3.41$, $p = .06$. There were no other significant main effects or interactions.

Tracking error. Increased tracking difficulty again increased tracking error ($M = 172.45$ vs. 316.50), $F(1, 11) = 675.66, p < .01$. Also, there was a marginally significant cost to tracking error when an auditory alert was presented, as compared with the visual alert ($M = 250.36$ vs. 238.60), $F(1, 11) = 3.59, p < .10$. This indicates that unlike the miss-prone system, for which multiple resource theory seemed to prevail, the FA-prone system seems to engender a modest preemptive effect with auditory alerts. There were no other significant interactions among alert modality, alarm type, or tracking difficulty on tracking performance.

Between-Experiment Comparisons

Response time. A final set of analyses examined the effects of alerting threshold by comparing the results of Experiment 1 with those of Experiment 2. In the following, we discuss only effects involving the experiment (i.e., threshold level) independent variable. Lowering the alert threshold had no overall impact on pilots' response time to CDTI conflicts (M = 9.49 vs. 8.86), $F(2, 22) = 0.22, p > .10$. However, as shown in Figure 5, alert threshold did interact with tracking stability, $F(2, 22) = 4.37, p < .05$, indicating that the FA-prone system increased RT, but only when the tracking was stable.

There was also a marginally significant interaction of alert threshold with alert type (binary vs. likelihood), $F(2, 22) = 3.42, p = .07$, indicating that the FA-prone system slowed RT, but only with likelihood alerts.

Sensitivity. Pilots were much less accurate in detecting CDTI conflicts with the miss-prone system (M = 1.89) than with the FA-prone system (M =3.38), F(2, 22) = 65.71, p < .01, even though the alerting systems themselves were nearly equally sensitive for both experiments (automation binary d'= 1.33 vs. 1.54 for Experiments 1 and 2, respectively; automation likelihood d'= approximately 1.55 and 1.52, respectively).

Tracking error. Tracking error was reduced with the FA-prone system of Experiment 2 (M = 244.7), as compared with the miss-prone system of Experiment 1 (M = 273.70), F(2, 22) = 3.24, p = .05. This difference is consistent with the theory that as automation misses decreased, reliance increased, leading to less monitoring of the raw data in the alerted domain and therefore more visual resources devoted to the concurrent (tracking) task. Alert threshold did not interact with tracking difficulty, F(2, 22) = 2.58, p > .10, or with alarm type, F(2, 22) = 2.24, p > .10. That alert threshold did not interact with alarm type suggests that the likelihood alert does not mitigate the impact of increased interruptions by FAs. As shown in Figure 6, there was an interaction between modality and alert threshold, F(2, 22) = 4.64, p < .05, such that the decrease in tracking error with the decreasing system miss rate was greater with visual alerts than with auditory alerts.

Finally, in each experiment, we assessed reliance by examining RT on trials when the automation missed (long RT → high reliance) and assessed compliance by examining RT on automation hit trials (short RT → high compliance). In each experiment, there was an interaction between trial type (auto-miss vs. auto-hit) and alert type (binary vs. likelihood), Experiment 1: F(1, 11) = 10.39, p < .01; Experiment 2: F(1, 11) = 11.07, p < .01.

The pattern of the interactions showed that in Experiment 1, the binary alert showed both low reliance and low compliance, whereas the likelihood alert showed a fainter trend in the opposite direction. In Experiment 2, the binary alert showed both high reliance and high compliance, whereas the likelihood alert showed a smaller trend in the opposite direction. Thus, with the binary alert, lowering the threshold (from Experiment 1 to Experiment 2) and slightly raising the alert sensitivity (in terms of d') produced a substantial increase in automation dependence (increase in both compliance and reliance). In contrast, with the likelihood alert the same threshold shift produced a slight decrease in dependence.

DISCUSSION

The current experiments set out to evaluate four hypotheses that address potential effects and interactions among several factors relevant to alert design in a multitask context, representative

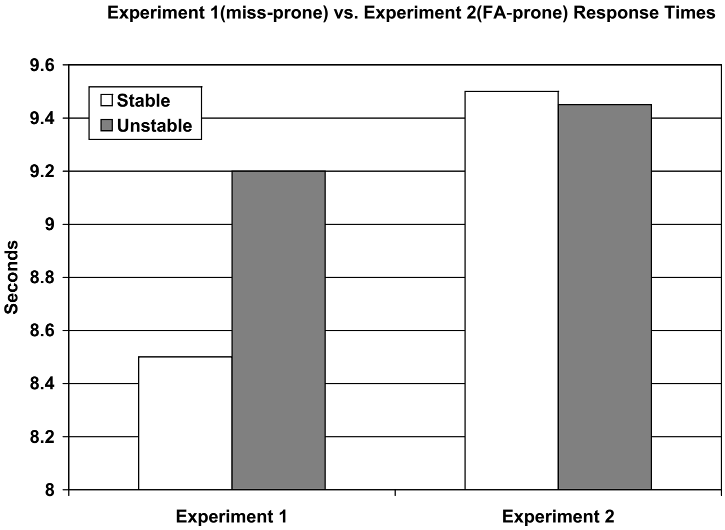

Figure 5. A significant two-way interaction for response time between alert threshold (Experiment 1, neutral, vs. Experiment 2, FA-prone) and tracking stability, such that increasing false alarm rate from Experiment 1 to Experiment 2 increased RT during stable, but not unstable, tracking.

of the information processing demands of flying. We examined four major hypotheses. First, we hypothesized that lowering the threshold to create the FA-prone system would reduce pilot compliance with the automated alert (Dixon & Wickens, 2006; Meyer, 2004) and consequently degrade conflict detection performance. This hypothesis was partially supported. Response times were indeed longer in Experiment 2, which had a higher FA rate than that in Experiment 1: a 1-s lengthening was observed when tracking was stable and when the likelihood alarm was employed, and RT was never shorter in Experiment 2.

We also assume that the increasing latency of switching attention from the tracking task to the alert task in Experiment 2 allowed more time for the conflict to develop and, thereby, was partially responsible for the increased detection accuracy in Experiment 2. Although it is possible that the increase in human sensitivity from Experiment 1 to Experiment 2 was also attributable to the slightly higher automation sensitivity ($d' = 1.50$ vs. 1.33 for binary; equal sensitivity for likelihood), we doubt that this small increase could explain the much larger increase in pilot sensitivity (3.38 vs. 1.89). Indeed, this large apparent benefit to sensitivity of a lower alert threshold level remains somewhat unexplained, but it does suggest that pilots in the FA-prone system were more optimally depending upon automation when it was correct.

Second, we hypothesized that the increased automation miss rate in Experiment 1 (as compared with Experiment 2) would reduce reliance, and hence degrade concurrent task (tracking) performance, as more attentional resources were diverted to monitoring the raw data of the traffic display. This effect was strongly supported (see Figure 6), as it has been in some other studies (e.g., Dixon & Wickens, 2006), but not in all (Dixon, Wickens, & McCarley, 2007; Levinthal & Wickens, 2005); the effect suggested that the addition of more automation FAs in Experiment 2 was not, here, particularly more disruptive to concurrent task performance. The analysis of RT to automation misses also indicated lower reliance in Experiment 2 than in Experiment 1, at least when the binary alert was used.

Third, we hypothesized that the auditory alert would improve the alerted (CDTI) task because of its attention-grabbing properties (Spence, 2001) and also because it would support parallel use of multiple perceptual resources (Wickens, 2002). This was indeed the case for conflict detection accuracy in both experiments and for RT in Experiment 2. (There was no effect of modality on RT in Experiment 1.) Concurrent flight control performance was unaffected by modality in Experiment 1, but it was marginally degraded by auditory alerts in Experiment 2.

In Experiment 1, the lack of a modality-based RT or tracking effect, coupled with increased

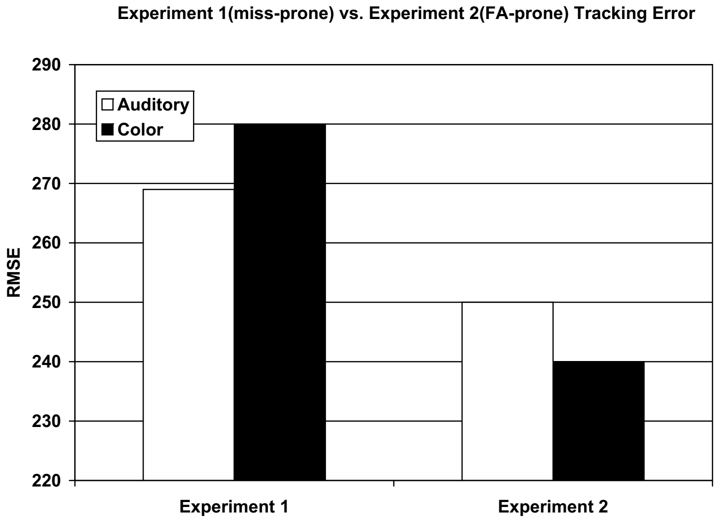

Figure 6. A significant interaction for tracking difficulty between alert threshold (Experiment 1, neutral, vs. Experiment 2, FA-prone) and alarm modality (auditory vs. visual), such that the decrease in tracking error with the decreased miss rate in Experiment 2 was greater for visual than for auditory alerts. RMSE = root mean square error.

accuracy for auditory alerts, seems consistent with the parallel processing features of multiple resource theory (Wickens & Hollands, 2000), even in this paradigm, in which noticing the visual alert did not require added scanning away from the tracking display. In Experiment 2, in contrast, faster and more accurate conflict detection, but degraded tracking performance caused by auditory alerts in the FA-prone system, suggests that auditory preemption (Iani & Wickens, 2007) prevailed with this system. This would imply that the greater number of total alerts brings about the auditory preemptive characteristic of those alerts.

Fourth, we hypothesized that the likelihood alert would aid conflict detection and particularly ongoing tracking (flight control) performance because it would allow pilots to distribute attention between the two tasks more optimally (Sorkin et al., 1988; St. John & Manes, 2002), essentially allowing a rapid switch only when the alerts were serious. However, we did not find evidence that the likelihood alert generally aided pilots in this way. In fact, we found costs associated with the likelihood alert in both the alerted domain (slower response times and, in Experiment 2, degraded accuracy for likelihood visual alerts) and the concurrent task domain (increased tracking error in Experiment 1).

CONCLUSION

In conclusion, the current results have both theoretical and applied implications. Regarding theoretical implications, the reliance-compliance distinction proposed by Meyer (2001, 2004) and reinforced by Dixon and Wickens (2006) can, in part, account for the results: as the threshold was shifted, both compliance-related cry-wolf effects on latency and reliance-related effects on concurrent tasks were observed. The great improvement in sensitivity as the threshold was shifted to produce the FA-prone system remains to be explained.

The data also suggest that both multiple resources and preemption appear to operate as mechanisms of task interference, with the relative contributions of one versus the other dictated by experimental factors – in particular, with auditory preemption emerging to dominate as the frequency of preemptive events (alerts) increased.

With regard to applications, the data support a low alert threshold that does indeed produce an asymmetry of automation errors favoring FAs over misses (and therefore reinforcing the typical cost matrix, in which the latter costs are more severe than the former). This was particularly true given the substantial improvement in both concurrent task performance and conflict detection accuracy in Experiment 2. However, we recommend caution in extending these benefits to more extreme low-threshold settings. In Experiment 2, the positive predictive value of the alert (likelihood that an alert will be true) was .70. In some systems with extremely low base-rate events, this value may be well below .50 (Getty et al., 1995; Krois, 1999), and in these cases, the escalating FA rate may amplify the otherwise minor cry-wolf effects observed in the current results.

The data also speak to the general benefit of auditory alerts; however, it is plausible to assume that their benefits may diminish as the frequency of false alerts increases substantially. Even here, we found the auditory preemption mechanism beginning to dominate the multiple resource benefit in Experiment 2.

We also observed that the status and benefits of the likelihood alert remain unclear. Its direct benefits, observed (Sorkin et al., 1988) and predicted (Woods, 1995) elsewhere, failed to emerge in either experiment. The only indirect benefit appeared to be manifest with the higher threshold setting in Experiment 1, in which the likelihood alert appeared to engender more dependence on automation (increased reliance and compliance), a characteristic that with imperfect automation is not altogether good. Clearly then, a considerable degree of further research is required to understand what circumstances, if any, will realize the proposed benefits of the likelihood alarm concept.

Finally, we note here some limitations of the current paradigm that partially limit its generalizability to cockpit alerting and warrant more research in more realistic flight simulations. First, the tracking task was a low-fidelity simulation of actual flight control and probably imposed greater visual demands (because of its high bandwidth) than would be typical. Second, the CDTI was somewhat larger than might be characteristic of some proposed CDTI designs, and hence visual time-sharing could be placed at more of a premium in the real cockpit. Finally, with regard to the traffic task, our conflict event rate was certainly higher than would be expected in a typical airspace, even one with great traffic density (e.g., in an unstructured terminal environment). These three factors certainly dictate the advisability of more research in more realistic flight simulations.

ACKNOWLEDGMENTS

This research was supported by a grant from the NASA Ames Research Center, NASA NAG-2-1535. The authors wish to thank the grant's scientific/technical monitor, Dr. David Foyle, and to thank Ronald Carbonari for software development.

REFERENCES

Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms.* Hillsdale, NJ: Erlbaum.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle flight control: A reliance-compliance model of automation dependence in high workload. *Human Factors, 48,* 474–486.

Dixon, S. R., Wickens, C. D., & McCarley, J. M. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses. *Human Factors, 49,* 564–572.

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied, 1,* 19–33.

Iani, C., & Wickens, C. D. (2007). Factors affecting task management in aviation. *Human Factors, 49,* 16–24.

Johnson, W. W., Battiste, V., & Bochow, S. H. (1999). A cockpit display designed to enable limited flight deck separation responsibility. Paper presented at the World Aviation Conference held October 19–21 in San Francisco.

Krois, P. (1999). *Alerting systems and how to address the lack of base rate information* (unpublished manuscript). Washington, DC: Federal Aviation Administration.

Kuchar, J., & Yang, L. (2000). A review of conflict detection and resolution modeling methods. *IEEE Transactions on Intelligent Transportation Systems, 1,* 179–189.

Latorella, K. A. (1996). Investigating interruptions – An example from the flightdeck. In *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting* (pp. 249–253). Santa Monica, CA: Human Factors and Ergonomics Society.

Levinthal, B. R., & Wickens, C. D. (2005). *Supervising two versus four UAVs with imperfect automation: A simulation experiment* (Tech. Rep. AFHD-05-24/MAAD-05-7). Savoy: University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Human Factors Division.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by human operators undermine trust in automation aids. *Human Factors, 48,* 241–256.

Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. *Human Factors, 45,* 281–295.

Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: effects of decision aid reliability on controller performance and mental workload. *Human Factors, 47,* 35–49.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors, 43,* 563–572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors, 46,* 196–204.

Parasuraman, R. (1987). Human-computer monitoring. *Human Factors, 29,* 695–706.

Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver-centred collision-warning systems. *Ergonomics, 40,* 390–399.

Pollack, I., & Madans, A. B. (1964). On the performance of a combination of detectors. *Human Factors, 6,* 523–541.

Schutte, P. C., & Trujillo, A. C. (1996). Flight crew task management in non-normal situations. In *Proceedings of the 40th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 244–248). Santa Monica, CA: Human Factors and Ergonomics Society.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors, 30*, 445–459.

Sorkin, R. D., & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human-Computer Interaction, 1,* 49–75.

Spence, C. (2001). Crossmodal attentional capture: A controversy resolved? *Advances in Psychology, 133,* 231–262.

St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting* (pp. 332–336). Santa Monica, CA: Human Factors and Ergonomics Society.

Swets, J. A. (1991). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47,* 522–532.

Thomas, L. C., & Rantanen, E. M. (2006). Human factors issues in implementation of advanced aviation technologies: A case of false alerts and cockpit displays of traffic information. *Theoretical Issues in Ergonomics Science, 7,* 501–523.

Thomas, L. C., & Wickens, C. D. (2005). *Effects of display dimensionality, conflict geometry, and time pressure on conflict detection and resolution performance using a cockpit display of traffic information* (Tech. Rep. AHFD-05-4/NASA-05-1/NASA NAG 2-1535). Savoy: University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Human Factors Division.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3,* 159–177.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science, *8,* 201–212.

Wickens, C. D., Dixon, S. R., & Ambinder, M. S. (2006). Workload and automation reliability in unmanned air vehicles. In N. J. Cooke, H. Pringle, H. Pedersen, & O. Connor (Eds.), *Advances in human performance and cognitive engineering research: Vol. 7. Human factors of remotely operated vehicles* (pp. 209–222). Amsterdam: Elsevier.

Wickens, C. D., Dixon, S. R., Goh, J., & Hammer, B. (2005). Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis. In R. Jensen (Ed.), *Proceedings of the 13th International Symposium on Aviation Psychology* (CD-ROM). Oklahoma City, OK: Federal Aviation Administration.

Wickens, C. D., Dixon, S. R., & Johnson, N. R. (2006). Imperfect diagnostic automation: An experimental examination of priorities and threshold settings. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 210–214). Santa Monica: CA: Human Factors and Ergonomics Society.

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W., & Talleur, D. A. (2003). Attentional models of multi-task pilot performance using advanced display technology. *Human Factors, 45,* 360–380.

Wickens, C. D., Helleberg, J., & Xu, X. (2002). Pilot maneuver choice and workload in free flight. *Human Factors, 44,* 171–188.

Wickens, C. D., & Hollands, J. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Wickens, C. D., & Liu, Y. (1988). Codes and modalities in multiple resources: A success and a qualification. *Human Factors, 30,* 599–616.

Wiegmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human Factors, 44,* 44–50.

Woods, D. D. (1995). The alarm problem and directed attention in dynamic fault management. *Ergonomics, 38,* 2371–2393.

Xu, X., Wickens, C. D., & Rantanen, E. M. (2007). Effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic information. *Ergonomics, 50,* 112–130.

Christopher D. Wickens is a senior scientist at Alion Science Corporation, Micro Analysis & Design Operations, Boulder, Colorado, and professor emeritus at the University of Illinois at Urbana-Champaign. He received his Ph.D. in psychology from the University of Michigan in 1974.

Angela M. Colcombe is a systems engineer with the Boeing Company, Seattle, Washington. She received her Ph.D. in psychology from the University of Illinois at Urbana-Champaign in 2006.

Date received: July 6, 2006
Date accepted: February 2, 2007