

INTEGRATING IMPERFECT AUTOMATED AIDS INTO A MULTI-TASK SITUATIONS

BY

ANGELA MARIE COLCOMBE

B.S., Northern Michigan University, 1995
A.M., University of Illinois at Urbana-Champaign, 2002

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2006

Urbana, Illinois

CERTIFICATE OF COMMITTEE APPROVAL

*University of Illinois at Urbana-Champaign
Graduate College*

December 12, 2005

We hereby recommend that the thesis by:

ANGELA MARIE COLCOMBE

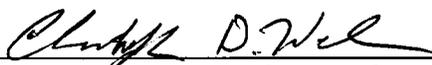
Entitled:

**INTEGRATING IMPERFECT AUTOMATED AIDS INTO MULTI-TASK
SITUATIONS**

Be accepted in partial fulfillment of the requirements for the degree of:

Doctor of Philosophy

Signatures:

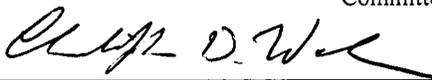


Director of Research

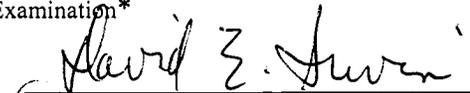


Head of Department

Committee on Final Examination*



Chairperson

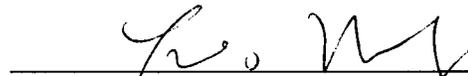


Committee Member

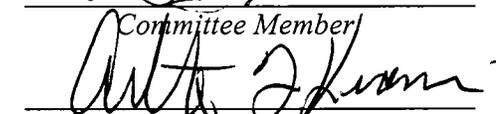
Committee Member



Committee Member



Committee Member



Committee Member

* Required for doctoral degree but not for master's degree

ABSTRACT

Automated aids are typically designed to help people monitor for particular events or situations during complex tasks. Automated aids can be useful by taking over a function once performed by the human thereby freeing up the person to do other tasks. However, automated aids may also introduce costs to task performance that are not readily apparent. In four experiments, we examined the impact of interruptions by an alarm embedded within a Cockpit Display of Traffic Information (CDTI) on two types of concurrent tasks; a compensatory tracking task and a working memory task. In addition, we examined how three important alarm characteristics; the informativeness (three vs. two stage alerts), the modality, and the threshold of the alert, affected both conflict detection and concurrent task performance. Automated aids with higher false alarm rates resulted in poorer concurrent task performance, as evidenced by higher tracking error and reduced working memory accuracy. Likelihood alerts did not mitigate costs associated with performance decrements due to the reduced alert threshold. Finally, auditory alerts tended to produce an auditory preemption effect, driving attention quickly to the alerted domain at the expense of concurrent task performance.

ACKNOWLEDGEMENTS

Many thanks to my academic advisor Dr. Chris Wickens for his tireless support and guidance on this project. Thanks to Dr. Art Kramer, Dr. Dave Irwin, Dr. Daniel Morrow, and Dr. Frances Wang for their insightful comments and for serving on my dissertation committee.

This is dedicated to my husband Stanley Colcombe, my daughter Parker Colcombe, my mother Sally Ewing, and my sister Therasa Healy for their loving support and encouragement.

TABLE OF CONTENTS

1. INTRODUCTION.....	1
1.1 The Benefits and Costs of Automated Alarm Systems	1
1.2 Theoretical Basis	2
2. AUTOMATED ALARM SYSTEMS	4
2.1 Alarm Threshold and Imperfect Alerting	4
2.2 The Cognitive States of Reliance and Compliance	6
2.3 Automated Alarm Systems in Multiple Task Environments.....	9
3. INTERRUPTIONS AND TASK INTERFERENCE	14
3.1 General Views on Interruption.....	14
3.2 Interruptability of Ongoing Tasks	14
3.3 Interrupting Task.....	17
4. MITIGATING ALARM DISRUPTION.....	21
5. CONFLICT DETECTION	23
6. SUMMARY	24
6.1 Purpose.....	24
6.2 Hypotheses	25
7. METHODS	28
7.1 Participants.....	28
7.2 Design	28
7.3 Procedure.....	29
7.4 Pilot Study.....	33
7.5 Dependent Measures.....	33
8. RESULTS: EXPERIMENT 1	34
8.1 Alerted Task Performance	34
8.2 Concurrent Task Performance	36
9. METHODS: EXPERIMENT 2.....	38
9.1 Participants.....	38
9.2 Design	38
10. RESULTS: EXPERIMENT 2.....	40
10.1 Alerted Task Performance	40
10.2 Concurrent Task Performance	42
11. RESULTS: EXPERIMENT 1 VS. EXPERIMENT 2.....	43
11.1 Alerted Task Performance	43
11.2 Concurrent Task Performance	45

12. METHODS: EXPERIMENT 3.....	48
12.1 Participants.....	48
12.2 Procedure.....	48
13. RESULTS: EXPERIMENT 3.....	50
13.1 Alerted Task Performance.....	50
13.2 Concurrent Task Performance.....	51
14. RESULTS: EXPERIMENT 2 VS. EXPERIMENT 3.....	54
14.1 Alerted Task Performance.....	54
15. METHODS: EXPERIMENT 4.....	56
15.1 Participants.....	56
15.2 Procedure.....	56
16. RESULTS: EXPERIMENT 4.....	58
16.1 Alerted Task Performance.....	58
16.2 Concurrent Task Performance.....	59
17. RESULTS: EXPERIMENT 3 VS. EXPERIMENT 4.....	60
18. RESULTS: EXPERIMENT 3 VS. EXPERIMENT 4.....	61
19. DISCUSSION.....	62
19.1 Interpretation of Results.....	62
19.2 Practical Implications.....	69
REFERENCES.....	71
CURRICULUM VITAE.....	77

1. INTRODUCTION

1.1 The Benefits and Costs of Automated Alarm Systems

Automated alarm systems are designed to detect and draw attention to some event in the environment. Alerting systems are most important in very complex, multitask, high stress, work environments such as aviation, industrial supervision, or even driving where human attention is at a premium. One benefit of automated alerting systems is that they can alleviate some of the humans' workload by taking over the task of monitoring for potentially unsafe conditions. This frees up the human operator's time and often visual resources which allows them to do other things. Alerts that serve as safety warnings are common in everyday life. For example, many people have an auditory alert that sounds when they forget to fasten the seatbelt in their car. Less familiar to most, are those alerts employed in very complex workstations, such as an aircraft cockpit, that may warn of a potential collision or equipment malfunction. Clearly the importance of an alert, the required action to restore safety, and the cost an alarm may introduce to ongoing tasks, can vary greatly. These variations of seriousness depend in part on the domain (e.g., aviation or office workspace), tasks at hand (flying an aircraft or writing a report), and the event indicated by the alert (deadly collision threat, or an e-mail alert). The complex interactions between automated alerting systems and the environments within which they are employed has presented a major challenge for cognitive science and the study of human-computer interaction.

Although it is true that technological advances in automation have allowed humans to do many more concurrent tasks than they could unaided, such systems have also introduced interruptions into already complex integrated workplaces. Such technology, when not designed for people's unchanging cognitive abilities, can lead to decrements of performance on ongoing primary tasks, due to the need of the human operator to disengage from the primary task and attend to alarms or notifications from the automated system (Woods, 1995; McFarlane & Latorella, 2001; Dixon and Wickens, in press). Literature from many domains has reported significant performance costs in interruption-laden, technologically advanced, work environments (McFarlane & Latorella, 2001; Gillie & Broadbent 1989; Bailey, Konstan & Carlis, 2003). Automated interruptions are particularly problematic when the technology has not been designed to compliment the human user's cognitive abilities in multi-tasking situations

(McFarlane & Latorella, 2002). Moreover, these costs are compounded with additional cognitive biases that result from the fact that in any meaningfully complex system, the automated diagnostic algorithms are imperfect (Meyer, 2001, 2004). And as such, the system will occasionally miss critical events (misses) that require user intervention as well as falsely identify non-critical events as those needing user intervention (false alarms). The nature and cost of these interruptions to ongoing task performance, including how the characteristics of the alarm notifications mediate these effects, must be addressed in order to minimize the cost of automation interruptions and maximize the benefits of intelligent and potentially beneficial systems.

Of interest here are issues related to imperfect automation within the specific domain of traffic conflict detection in aviation. Cockpit Displays of Traffic Information (CDTIs) are a developing technology that cue the pilot to potential air traffic collision threats. Designers tend to set the criteria threshold of systems designed to detect potentially catastrophic events in a way that minimizes misses. Therefore, systems like the CDTI tend to be false-alarm prone in terms of automation errors. This fact is a crucial issue to this dissertation work because it is thought to influence pilots' psychological state, cognitive strategies, and the impact of mitigating characteristics of the technology itself (Wickens, Helleberg & Xu, 2002).

1.2 Theoretical Basis

In this study, we address how interruptions, characteristic of imperfect alerting, affect ongoing task performance. We are particularly interested in how parameters of the ongoing task make them more or less interruptible from an alert, and in turn, how the characteristics of the alert mediate the costs of interruptions to ongoing tasks. Three somewhat different bodies of literature inform this project. 1) Literature on multi-task performance suggests that performance decrements will ensue as a person attempts two or more tasks at once depending on the tasks' difficulty, priority, and modality, as well as the strategy adopted by the person (e.g., Wickens & Holland, 2000). 2) Literature on interruptions provide more specific information about the costs associated with the timing, urgency, frequency and modality of interruptions as well as scheduling of interrupting and ongoing tasks (e.g., McFarlane & Latorella 2002; Gille & Broadbent 1989; Woods, 1995; Monk 2004). 3) Finally, literature on imperfect automated alerts

provides guidelines for predicting how a human will interact with an imperfect alarm system that necessarily introduces interruptions within a complex multi-task environment, given cognitive biases that result from different kinds of automation errors (e.g., Dixon & Wickens, in press; Maltz & Meyer, 2001; Iani & Wickens 2004). It is these three bodies of literature that have guided our topic of interest. Literature on automated alarm systems, dual-task interference, and interruptions will be discussed next to provide a basis and rationale for our experiments.

2. AUTOMATED ALARM SYSTEMS

2.1 Alarm Threshold and Imperfect Alerting

Alerting systems are extremely useful in multi-task, high-stress work environments, such as aviation, in that they can direct human attention to a critical event that may require immediate attention (Parasuraman, Sheridan & Wickens 2000, Pritchett, 2001). However, automated alarm systems tend to be imperfect, particularly to the extent that they forecast future events based on the state of the world at a given point in time. This involves a degree of uncertainty because factors contributing to a given state can be ambiguous or change as time goes by. For example, in aviation, automated conflict detection is imperfect due to factors such as wind shifts, turbulence, or intentional changes executed by the pilot (Kuchar, 2001; Thomas et al., 2003). In medicine, diagnostic alerts tend to be imperfect because the data they are based on is often ambiguous or fuzzy (Swets, 1992). Because of this uncertainty, automated system decisions are often discussed in terms of signal detection theory, with error rates associated with false-alarms and misses being emphasized. The potential outcome of an automated decision (hit, miss, correct rejection, false alarm), as framed in signal detection theory, depends on the sensitivity of the aid, or its' ability to detect a given event (signal) in the environment (noise), and the threshold of the alert, or its willingness to indicate an event will occur given available evidence at a particular time. The factors affecting automation misses and false-alarm rates will now be discussed in more detail.

One factor that influences the error rates of an alarm is its sensitivity as dictated by the quality of the algorithm that is utilized to discriminate signal from noise and the quality of the data to be detected. The algorithm is constrained by what is technologically possible, and the choices of the designer. In addition to not detecting all hazardous events, the automation may also fail to detect all collision threats, such as a blimp, terrain, or a building, if it was not designed to do so. The prediction made by the automation concerning a potential collision threat is further complicated by the fact that an aircraft on a certain path may change flight parameters at any point so that the degree to which they pose a threat can change across time. That is, the further ahead the prediction is being made, the more error-prone the system will be. Of course with little to no "look ahead time" an alert system could be 100% accurate. However, if alerts

sounded just before a dangerous event occurred (e.g., a collision), the pilot would have little time to maneuver to safety. It is for this reason it is desirable to create systems that can predict events sufficiently ahead of when they would occur to allow time for the pilot to restore safety.

The fact that an automated aid cannot possibly detect all hazardous events, or even all events it is designed to detect, implies a shared responsibility between the automation and the user. Because of this, the pilot must take some responsibility in monitoring the environment for such events (Sorkin & Woods, 1988). So, while the automation will aid the user in detecting collision hazards, it should not be completely relied on to the exclusion human monitoring of possible collision events. Unfortunately, it has been empirically demonstrated that humans do sometimes rely completely on the automated aid without engaging in the necessary monitoring themselves (Thomas & Wickens, 2004), especially with systems that have been designed to minimize misses. This phenomenon will be addressed more thoroughly later. Now we turn to a second factor that affects automated decision error-rates, the alert threshold.

While the sensitivity of an automated aid is typically static, the threshold, or response criterion (Beta) can be manipulated to minimize one type of automation error (e.g., misses) over another (e.g., false alarms). The trade-off between these costs can guide the decision regarding the placement of the threshold for the system. One factor affecting the placement of automated thresholds is the probability of the event that it is designed to detect, sometimes referred to as the “base rate” (Parasuraman, Hancock, & Obofinbaba, 1997). If the probability of an event is high, a relatively low threshold setting will result in an optimal balance of errors and detection rates, a minimization of both types of errors. In contrast, if the probability of an event is low, as is typically the case in the real world where “normal” conditions prevail over abnormal ones, a higher threshold setting will result in more optimal automated decisions. However, a second extremely important factor to consider when choosing the threshold for an automated systems is the cost associated with each of its decision outcome errors. For example, in aviation, the probability of another aircraft invading ones own airspace, a conflict, is quite rare, but it clearly presents a significant and potentially dangerous threat. This type of event should not be missed.

One way to minimize misses is to set the detection threshold for the system quite low so that the system liberally detects events. While this seems to be the safest course of action, such a

low threshold setting, based on the combination of a very low base rate of events and a high cost of missing such events, leads to numerous false-alarms (Krois, 2001; Parasuraman et al., 1997). The implications of missing a dangerous event are clear. In contrast, the costs associated with high false-alarm rates are more elusive and have been the source of some recent research in the literature on automation and interruptions (e.g., McFarlane & Latorella, 2001; Dixon & Wickens, in press). Systems that are miss-prone or false-alarm prone, engender different levels of trust and dependence in human users. When the threshold is set low, and consequently the system makes few miss errors, ironically performance costs due to human over-reliance on the system seem to ensue (e.g., Bainbridge, 1983; Parasuraman & Molloy, 1996; Dixon & Wickens, in press; Maltz & Shinar, 2003). This leads to faulty detection by the human of the now very rare events that automation also misses. That same low threshold also results in a false alarm-prone system that leads to reduced compliance with the alert, the so called “cry-wolf” effect (Breznits, 1983). The costs related to miss-prone and false-alarm prone systems map onto two cognitive states known as **reliance** and **compliance**, which in turn result in particular behaviors that can be detrimental to performance (Meyer, 2001, 2004; Maltz & Shinar, 2003; Dixon & Wickens, in press, Wickens, Dixon, Goh, & Hammer, 2005; Wickens & Dixon, 2006). The two cognitive states which are affected by both miss-prone and false-alarm prone systems are now discussed in terms of the cost each produces.

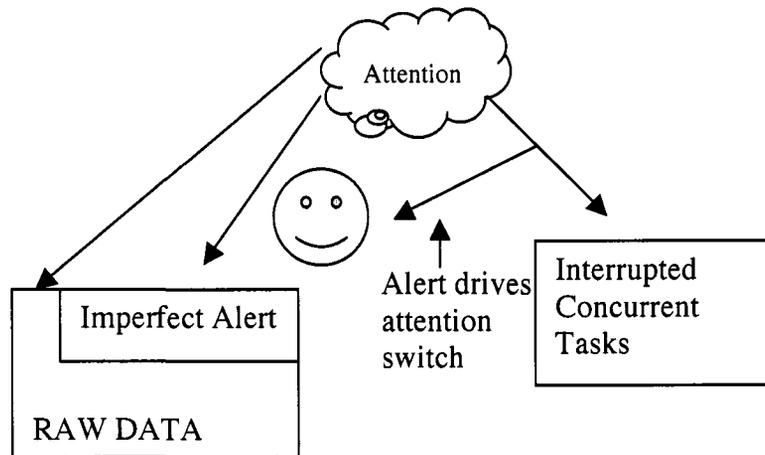
2.2 The Cognitive States of Reliance and Compliance

Reliance is a cognitive state that occurs with reliable automated systems that have a very low miss rate (Meyer, 2004). Very reliable detection systems with a low miss rate engender a high level of trust from users, and the user feels free to focus attention on other tasks when the alert does not indicate a potential threat (e.g., Bainbridge, 1983; Parasuraman, Molloy & Singh, 1993; Lee & Moray, 1994; Parasuraman & Riley, 1997). During an “all safe” state, the human user relies on the automated aid to alert him or her if the state of the world becomes dangerous. One potential problem associated with reliance, as mentioned above, is when the human relies exclusively on the automated aid and neglects to check the “raw data” periodically to assure that the situation is indeed safe. This can lead to the rare automation miss being undetected by the human as well. This phenomenon is referred to in the literature as **complacency** (Parasuraman et al., 1993; Parasuraman & Riley, 1997). Complacency is most detrimental the first time an

automation miss occurs (e.g., Molloy & Parasuraman, 1996; Davison & Wickens, 2001; Yeh et al., 2003). Because the human is completely relying on the aid before the first miss, it is very unlikely that the human will detect the event, and therefore the likelihood of a serious accident increases (Wickens 2002). Nevertheless, some reliance effects are observed on subsequent automated misses as well (Wickens & Dixon, 2003; Yeh et al., 2003; Wickens et al., 2005).

On the other hand, as the miss rate increases, reliance decreases. This has consequences for the concurrent tasks carried out along with automation monitoring as depicted schematically in Figure 1. Decreased reliance means that the human will divert some resources from concurrent tasks to monitoring the raw data underlying the automated aid to ensure that no misses occur. Of course, while this means that the human will be less likely to miss detecting important events even if the automation does (i.e., reduced complacency effect), it also means that performance associated with the ongoing tasks can suffer because resources once devoted to it are now being directed toward monitoring the alerted domain (Wickens & Dixon, 2003; Wickens et al., 2005). And of course, when automation has a sufficiently high miss rates, it can cease helping and become a liability (Dixon & Wickens 2004; Maltz & Shinar, 2003).

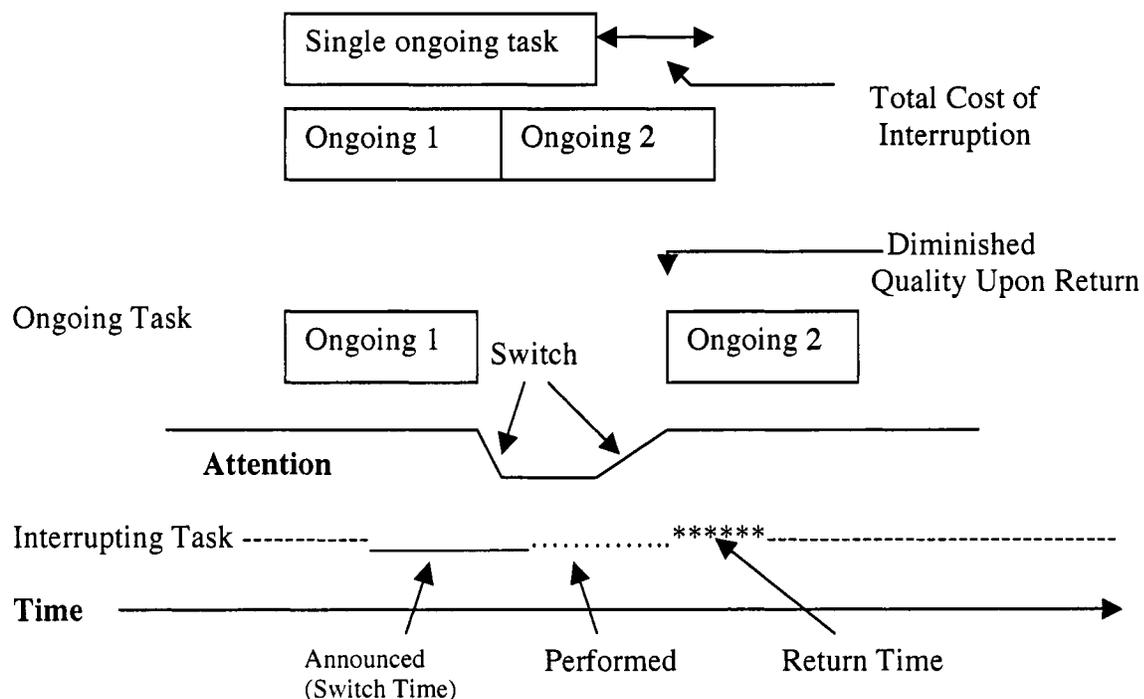
Figure 1: An Illustration of how attention is distributed between ongoing concurrent tasks and the automated domain.



The other cognitive state associated with automated alerting systems is **compliance** (Meyer 2001, Dixon & Wickens 2003; Maltz & Shinar, 2003; Breznitz, 1983), which describes how rapidly and accurately the human responds to the alert when it sounds. Humans tend to take alarm systems that have very low false-alarm rates quite seriously, and therefore disengage from ongoing tasks to comply with the alert rapidly (Wickens 2004). This rapid compliance is often

good in terms of performance in responding to the aid, especially in time constrained scenarios. However, there can be a cost to blind compliance when the system makes even an occasional false-alarm error. In this case the user may initiate an action, such as changing flight parameters if an alert indicates another aircraft is on a collision course with one's own, without even verifying the accuracy of the alert. By definition, alerts impose interruptions onto the concurrent tasks that the pilot must do. As shown in Figure 2, even if the pilot takes no action other than simply to verify the authenticity of an alert, performance on the concurrent task may suffer when attention is returned to it because of the need to reorient (McFarlane & Latorella, 2001; Bailey et al., 2003; Adamczyk & Bailey 2003). This can be especially wasteful of attentional and cognitive resources if the alert was false to begin with. The negative impact of abruptly dropping an ongoing task to attend to an alarm depends on a number of factors that will be discussed in more detail later in this paper.

Figure 2: A schematic of the costs associated with interrupting an ongoing task.



While a low false-alarm rate engenders a high level of compliance and consequently, a rapid attentional shift toward the automated alarm, a false-alarm prone system will lead to reduced compliance, and a slower response to the alarm. As noted above, false alarm prone

systems are common in situations where it is desirable to minimize automation misses, so the designer will choose to lower the detection threshold (Wickens & Hollands, 2000; Swets 1992; Dixon & Wickens, in press). With a false-alarm prone system, compliance decreases, manifested in slower switching from ongoing tasks to deal with the alarm (Wickens et al., 2005) or even ignoring the alarm altogether, the “cry wolf effect” (Wickens, 2004; Breznitz, 1983). For example, Cotte, Meyer, and Coughlin (2001) found that drivers were less likely to comply with a forward collision warning system when it exhibited a high rate of false alarms.

False-alarms prone systems have been shown to be more detrimental to performance compared to miss-prone systems (Meyer 2001; Dixon & Wickens, in press; Maltz & Shinar, 2003). Maltz and Shinar (2003) found that increasing the automated false alarm rate hurt subjects’ performance on a target detection task, but increasing automated miss rate did not have an adverse effect. However, these investigators manipulated miss and false alarm rates within a single task scenario, which does not speak to performance costs associated with concurrent tasks. In multiple task environments, where alarms are most likely to be used, the effects on performance due to high false-alarm rates need to be considered for both the automated task and concurrent tasks (Dixon & Wickens 2004; in press). With less attention devoted to the alerted domain (due to the “cry-wolf” effect), users are less likely to detect automation misses, which could result in a catastrophic event. In addition, false-alarms impose unnecessary interruptions on concurrent task that may impose a cost to ongoing task performance.

2.3 Automated Alarm Systems in Multiple Task Environments

The concepts of reliance and compliance help to describe and predict how humans will interact with alarm systems. However, these concepts also illuminate how performance will be affected for ongoing concurrent tasks as depicted in Figure 1. Studies that measure responses to both the alerting task and ongoing tasks at different automation detection thresholds are especially important because they provide a comprehensive analysis of how an alerting system will impact performance in the types of multiple task environments they are most likely to be used. Table 1 illustrates the potential consequences for the automated *and* concurrent tasks as a function of the alert threshold setting. In the following section, literature is reviewed that has investigated automated alarms with high and low threshold settings in dual task situations.

Table 1: The consequences of different error rates as a function of threshold setting.

	High Threshold (Conservative)	Low Threshold (Liberal)
False Alarm Rate (Compliance)	<ul style="list-style-type: none"> -Low FA rate -Automated task improves -Rapid switch from concurrent task may negatively impact concurrent task 	<ul style="list-style-type: none"> -High FA rate -Slower switch to alert leads to decreased alert performance but improved concurrent task performance -Unnecessary interruptions lead to decreased concurrent task performance
Miss Rate (Reliance)	<ul style="list-style-type: none"> High Miss rate -more attention to raw data -concurrent task suffers 	<ul style="list-style-type: none"> -Low Miss rate -Improves detection rate -But complacency can lead to human miss -Less attention to raw data leads to improved concurrent task performance

Four studies (Dixon & Wickens, 2003; Dixon & Wickens in press; Wickens et al., 2005; Wickens, Dixon, & Johnson, 2005) are most relevant to this dissertation because they address the impact of the alert threshold on performance of both the automated task and concurrent tasks, and discuss the results in terms of reliance and compliance. Dixon & Wickens (2003; Dixon & Wickens, in press) compared performance on both an automated event monitoring task, and concurrent tasks, as the number automated FA and misses on the alert was varied, in an effort to independently manipulate compliance and reliance. As anticipated, they found that increasing automation miss rate (thereby decreasing reliance) degraded concurrent task performance, but improved the ability to detect the (now more frequent) automated misses (e.g., reduced “complacency”). They also found increasing automated FA rate decreased the speed of responding to events of the monitoring task (particularly in high workload), the so called “cry wolf effect” (Breznitz 1983). However, somewhat surprisingly, this FA induced reduction in compliance also lead to a parallel effect of increasing reliance, as if the pilots assumed that the more false alarm-prone system would, therefore be less miss-prone. Wickens, Dixon, Goh & Hammer (2005) used the same experimental simulation as Dixon & Wickens (in press) with a direct measure of attention. The authors (Wickens et al., 2005) found that an increase in automation miss rate led to longer (eye) dwell times on the raw data. In addition, they found that increases in automation false-alarm rate led to increased visual attention switch time (to the system monitoring display) and longer dwell times (to verify the authenticity of the alert). In other words, the authors (Wickens et al., 2005) found direct attentional costs to imperfect automation.

Bustamante, Anderson and Bliss (2004) examined how alarm threshold and task complexity affected performance in a dual-task situation. Participants in Bustamante et al.'s (2004) experiment performed a compensatory tracking and a resource management task, but were also responsible for system monitoring as a secondary task. The threshold of the alarm system was manipulated as was the difficulty of the tracking task. Unfortunately, the only dependent measure in the experiment was subjects' mean reaction time to the system monitoring task. The automated aid improved performance on the monitoring task (speeded reaction time) regardless of its' threshold setting when combined with the easier tracking task. However, during the difficult tracking task, monitoring improvement was only obtained when the threshold of the alarm was set at a low or medium level, but not when it was set at a high level. This is counter-intuitive because in the condition where the alarm threshold was set low, the false alarm rate was higher compared to the condition where the alarm threshold was set high. One might predict, under this circumstance, that the higher false alarm rate would reduce compliance with the automated task and therefore slow reaction time to the monitoring task. In fact, the opposite occurred. Response times were faster to the monitoring task when the false alarm rate was higher and slower in the condition when the false alarm rate was lower.

Lee, McGhee, Brown, & Reyes (2002) examined the effectiveness of early compared to late collision warnings for distracted drivers. Early warnings systems tend to be false-alarm prone compared to late warning systems because changes in the environment are more likely at longer time intervals. The authors varied initial velocity (35mph vs. 55 mph), severity of lead vehicle deceleration (low or high), and the timing (and therefore implicitly, the detection threshold) of the warning (early or late). A secondary task was introduced intermittently that, when engaged, resulted in an imminent collision situation at the very time the driver was distracted with the secondary task.

The authors found that early warnings resulted in 80% fewer collisions and reduced accident severity by 96% compared to the no warning condition. Late warnings also helped substantially compared to baseline, but much less so compared to early warnings. The findings of this experiment imply that a false alarm prone system (early warning) is better in terms of supporting performance compared to a miss prone system (late warning). Unfortunately, the authors of this experiment only included dependent measures related to responding to the alert

(such as brake speed), not to ongoing task performance (such as lane keeping). So, while it seems clear that the alert system will help a person respond to the alerted situation well, it is unclear what the impact of the system is on other driving tasks.

Thus, there are not many research findings that can be used to determine the impact of varying the automated false alarm rate and miss rate on both ongoing and automated task performance. The examples from the literature summarized above suggest that the relationship between the effects of alerting cues on response to automated alerts and ongoing task performance is complex. There appears to be only three studies that have examined alert response and concurrent task performance related to various automated false alarm rates (Dixon & Wickens, 2003, Dixon & Wickens, in press; Wickens, Dixon, Goh, & Hammer, 2005). Only two additional studies manipulate automated threshold settings under dual task conditions, but these studies unfortunately do not measure the impact of the alert on concurrent task performance (Bustamante et al., 2004; Lee et al., 2002). Within these five studies, there are inconsistencies that should be addressed.

One issue that remains unclear is exactly how false alarm prone systems affect overall task performance. On the one hand false alarm prone systems may improve concurrent task performance because the operator assumes “more false alarms means fewer misses” (which will be true to the extent that the increasing false-alarm rate is caused by a designer’s alert threshold adjustment), and this attitude engenders greater reliance. On the other hand, concurrent task performance might be improved because the high false alarm rate leads users to ignore the alerts altogether – the cry wolf effect (Breznitz, 1983). Alternatively, false alarm prone systems may **degrade**, rather than improve concurrent task performance (a) because of the high workload imposed by the added false alerts (Bustamante et al., 2004) or (b) because the salient false-alerts signal a general unreliability of the automation and thereby degrade reliance as well. More experiments can address this issue by including more dependent measures related to compliance and concurrent task performance. This seems particularly important for systems, such as CDTIs, that are designed to predict future events that may be especially hazardous because common sense might dictate a low alert threshold in order to minimize automation misses (Rantanen, Wickens, Thomas, & Xu, 2004). However, much still remains to be examined in regard to the

effects of specific alerting characteristics on ongoing primary tasks before this low alert threshold becomes a standard.

3. INTERRUPTION AND TASK INTERFERENCE

3.1 General Views on Interruption

Automated systems typically work on a delegated task in the background while a human is involved in other ongoing tasks. This is a clear advantage of automation in that it alleviates at least one task from the human and thereby allows them to concentrate on other things. While automated systems have the potential to support human performance in multi-task settings in this way, they also necessarily introduce interruptions (McFarlane & Latorella, 2002). These interruptions can be substantially increased in a situation where a low base-rate event is to be detected that has potentially deadly consequences if missed. In this case, as noted above, the threshold of the alarm is likely to be set low to minimize misses, but will thereby increase the number of false alarms (Parasuraman, Hancock, & Olofinboba, 1997; Krois, 1999). Interruptions due to a false alarm are not only annoying (Bliss, 2004; Maltz & Shinar, 2003), but interruptions have been shown to play an important role in ongoing task performance (Bailey et al., 2003; McFarlane & Latorella, 2002; Monk, 2004). Mainly, these costs tend to be associated with having to switch attention to another task and then back to resume the ongoing task. Figure 2 illustrates the time costs associated with switching attention in this manner.

3.2 Interruptability of Ongoing Tasks

Several factors play a role in predicting the disruptive effects of interruptions on ongoing concurrent tasks. One factor is the degree to which the ongoing task itself is vulnerable to performance decrements when it is interrupted. Some tasks are delayed due to interruptions but are not otherwise affected. For example, following a checklist should be immune to performance decrements due to interruptions as long as one's place is marked when the checklist is left. However, performance on complex tasks that rely heavily on working memory are likely to suffer when an interruption occurs. This is because information in working memory must be retained while one is on leave from this type of task so that the task can be resumed seamlessly once the interruption has been dealt with. Bailey, Konstan and Carlis (2003) found that tasks with a higher memory load suffered more from interruptions compared to tasks with a lower memory load. Evidence for this was, in part, due to the longer time spent on memory dependent tasks when they were interrupted compared to when they weren't. The logic is that memory

dependent tasks take longer to reorient to after an interruption compared to tasks that do not rely heavily on memory.

Due to limits of short term memory, interruptions may be costly to ongoing task performance if the necessary information cannot be rehearsed due to the demands of the interrupting task (Monk, Boehm-Davis, & Trafton, 2002). If the interrupting task prohibits rehearsal of ongoing task information, some of the information will probably have to be re-processed once the task is resumed, or in the extreme, the task may have to be started over (McFarlane, 1987; Miller, 2002) example Monk et al. (2002) found that the instruction to rehearse ongoing task information during an interruption was useless when the interrupting task itself was demanding in terms of working memory. We hypothesize that auditory working memory tasks will be more demanding compared to visual tasks because the visual task will enable the person to reorient but if auditory information is lost, it cannot be regained. Therefore we predict that an auditory working memory task (such as an ATC communication task) will be less interruptible compared to visual tasks (such as a data-link communication task). The high working memory demand of the auditory ongoing task may delay the switch to the alerted (interrupting) task if participants try to preserve the working memory task performance (Latorella, 1996). This may show up as improved performance on the auditory working memory task (compared to an equivalent task visually rendered), but an increased delay in switching attention to the interrupting task.

Another type of task that does not rely on working memory, but may be vulnerable to performance decrements due to interruptions, are those tasks that evolve, or change state, on their own when neglected. For example, a car would likely drift if one were to momentarily stop steering it. Iani and Wickens (2004) found slight flight path deviations after pilots diverted their attention to deal with the weather, which indicated that people were unable to maintain perfect tracking while attending to an interrupting task. The degree to which performance on this type of task would be susceptible to decrements due to interruptions will likely be related somewhat to the bandwidth of the task but particularly to the stability of the system being tracked. That is, systems that require frequent inputs from the human in order to remain stable or inherently produce positive feedback, would be more susceptible to performance decrements when they are interrupted compared to tasks that would require less frequent human intervention or are

inherently self-correcting (Wickens & Hollands, 2000). We hypothesize that the more vulnerable a task is to interruptions through working memory demands or instability, the less willing an operator will be to leave it when an alert occurs, and paradoxically, the more impervious the task will be to interruptions. In short, the human will perform some adaptive compensatory action on the ongoing task that will protect it from the disruptive effects of the alert.

Within any type of task, the degree to which one is **engaged** in that task may also be a factor in determining how interruptible it is. Engagement has not been operationally well defined, but includes things such as interest level, focus, and workload. Anecdotally, most people can describe situations when they've been so engrossed in a task they've failed to notice something that would normally interrupt them. In fact, some research has found that engaging tasks can be attentional sinkholes (Moray & Rotenberg, 1989). For example Moray & Rotenberg (1989) found that when subjects were engrossed in handling a system fault, devoting all attention to one display, they failed to notice other sources of information that indicated another fault had occurred. In fact, many aircraft accidents have been associated with pilot failure to shed tasks to attend to those that presented more imminent concerns for safety (Funk, 1991; Chao, Madhavan, & Funk, 1996; Wiener, 1977). Certain types of engaging or "compelling" displays may be less likely to support operator's switching attention to notice unusual events (Wickens, 2004; Thomas & Wickens, 2004).

Based on previous studies (Moray & Rotenberg, 1989; Wickens, 2004; Thomas & Wickens, 2004), Iani and Wickens (2004) hypothesized that participants in a flight simulation experiment would be less willing to switch attention from a compelling 3D immersive flight display compared to a less compelling one. That is, the more engaging display was expected to be more immune to the effects of interruption from a weather cue. Contrary to the prediction, the compelling display did not lead to reluctance to switch to the weather information, and in fact weather change detection was better for participants who used the compelling, but also more effective, display. In addition, flight path deviations after the weather change (interrupting task) were more severe for those using the compelling display compared to those using the baseline display. The authors suggest that these effects were likely due to the decreased workload afforded by the immersive display. The immersive display reduced the workload of flight path tracking and thereby freed attentional resources that could be diverted to the interrupting weather

change task. This led to better detection rates for the group who used the immersive display, but these distractions led to more overall tracking errors compared to the group that used a less compelling but more challenging baseline display and did not notice the weather changes as much. The findings of this study indicate that the degree of engagement in a task may interact with other factors, such as workload, in predicting how interruptible a given task will be.

3.3 Interrupting Task

In addition to the type and engagement level of ongoing tasks, characteristics of the interrupting task may also play a role in facilitating, or ameliorating the degree of disruption imposed by interruptions to ongoing task performance. In general, announced tasks, signaled by an external stimuli (auditory or visual cue), are more disruptive than unannounced tasks that rely on some internal decision on the part of the human to engage in them. The salience of an announced task will determine how disrupting it is in terms of ongoing task performance, with auditory signals typically being more salient and attention grabbing compared to visual cues (Latorella, 1996, Spence, 2001, Banbury et al., 2001; Wickens, Webb, & Fracker, 1987; Wickens & Hollands 2004; Iani & Wickens, in press). The goal for the alert system designer is to strike a balance between the attentional grabbing properties of an alert and its disrupting effects on ongoing task performance. Of course the context will probably play an extensive role in terms of choosing an appropriate level of cue salience that will achieve this balance, and will depend on how urgently the interrupting task must be attended to (Obermayer & Nugent, 1997).

The literature is somewhat mixed regarding performance tradeoffs for salient cues (auditory) compared to less salient cues (visual only). In a basic attentional capture paradigm, Spence (2001) found that discrete auditory stimuli will capture visual attention as well as auditory attention and will do so better than discrete visual stimuli. This is a benefit when it is desirable to switch all attentional resources to the cued (i.e., interrupting) task. However, the use of an auditory cue might lead to costs in terms of the ongoing task. Wickens & Liu (1988) suggest that the attentional grabbing properties of a discrete auditory cue will preempt an ongoing visual task, leading to improved performance on the interrupting (auditory) task, but at the expense of the visual ongoing task. Wickens, Dixon, & Seppelt (2001; 2005) observed this auditory preemption take place. However, this (auditory preemption) effect may not be

pronounced or observable in all situations. For example Iani and Wickens (2004) found that a salient auditory cue supported better weather change detection compared to a visual cue, but did so without imposing a cost to the ongoing visual tracking task. Also, Ho, Nikolic & Sarter (2001) found that disruptions imposed by an auditory interrupting task was less than that imposed by a visual task as reflected by performance of a visually demanding ongoing ATC task.

This finding may instead reflect the benefits of cross-modal presentation, consistent with the view of attention as multiple resources (Wickens, Goh, Helleberg, Horrey, & Talleur, 2003). Here the use of an auditory modality for an interrupting task, by using separate perceptual resources to the ongoing visual task, may improve performance on that ongoing task, as well as performance of the interrupting task, an effect that is known to amplify as visual separation between tasks is increased..

One factor that may contribute to whether cue salience affects ongoing task performance is the degree it disturbs memory processes. Banbury, Macken, Tremblay and Jones (2001) found that discrete auditory stimuli, such as those used with alarms, tend to corrupt memory processes more than visual cues. This would likely mean that returning to an ongoing task would be more difficult after an auditory interruption compared to a visual one, especially when working memory load is high. Helmick-Rich, Burke, Gilad, & Hancock (2004) found that people were less likely to comply with a visual cue compared to an auditory cue regardless of memory load. But, when working memory demand was high, a more salient (auditory) cue was necessary to generate compliance with the interrupting task. It would seem that while auditory cues are especially disruptive to ongoing memory tasks, they are sometimes needed in order to divert attention away from such tasks, especially when work load is high. This literature supports our hypothesis that auditory cues will lead to a rapid switch away from ongoing task to the alerted task, but this rapid switch will generate a cost to ongoing task performance.

Along with cue salience, the frequency, and timing of the interrupting task has been suggested to affect how disruptive it is (Monk, 2004; McFarlane & Latorella, 2001; Bailey et al., 2003). However, the results concerning these factors have been somewhat mixed. While it seems intuitive that more frequent interruptions would lead to worse concurrent task performance, some authors have found the opposite to be true (Monk 2004). Presumably more frequent interruptions

can lead to more stringent and effective task scheduling which leads to improved ongoing task performance in terms of efficiency (time to complete the task) and accuracy (Monk, 2004). This finding makes predicting the impact of real-world automated systems, such as conflict detection systems, on ongoing task performance more difficult. As mentioned earlier, the increased false-alarm rates of these types of systems can lead to reduced compliance, a negative effect associated with the interrupting (automated task). However, the increase in false-alarms may not necessarily have a negative impact on concurrent task performance if more frequent interruptions lead to better task management, or tend to be ignored.

In addition to the frequency of interruptions, the timing of interruptions may be important as well. Bailey and colleagues (2003) found that interrupting a task is best when subtasks have been completed but the next subtask has not been started. McFarlane & Latorella (2001; McFarlane, 2002) have argued that the timing of interruptions is an important factor in preserving particular subsets of task performance. That is, an aircraft collision conflict can occur at any time and is inherently urgent, and therefore the associated conflict detection alerting system would be forced to impose an immediate interruption. However, the point that interruptions will be more damaging if they occur at some points in an ongoing task than others is a good one, and has been confirmed by other experiments (Monk et al., 2002). Knowing that not all times are equal in terms of interruptions allows one to predict those times during a task that an interruption will be most damaging, and also implies that those are the times that some mitigating tools might be most useful. So, although conflict detection alerts will introduce immediate interruptions, the system can be designed to mitigate the effects of untimely interruptions by capitalizing on certain alert characteristics. This will be discussed in the next section that addresses specific alarm characteristics that may mitigate the disruptive impact of interruptions.

Overall, it would seem that the degree of disruption due to interrupting tasks is a complicated relationship between the characteristics of the ongoing task, those of the interrupting task, and the strategies and skills of the human. In general, it would appear that auditory alerts are more interrupting (attention capturing) to ongoing tasks than visual alerts, but despite this, they don't always impose a cost to ongoing tasks (Spence, 2000, Latorella, 1996; Iani & Wickens, 2004). More research is needed to address specific situations in which more salient

(auditory) alerts support performance and those situations in which such alerts are too costly to be beneficial. In addition, ongoing tasks are more disturbed when task demands, and in particular working memory demands, are high, which would be more likely if these tasks were entirely dependent on auditory information. Some research has indicated that interruptions do not always lead to mistakes (Lee, 1992), and that the key to minimizing errors due to interruption lies in automation design (McFarlane & Latorella, 2002). While there is no established set of guidelines to solve the problem of interruption induced errors yet (McFarlane & Latorella, 2002), it would seem that one part of the puzzle lies in understanding how alert characteristics affect ongoing task performance (Woods, 1995; Sorkin, Kantowitz & Kantowitz, 1988).

4.0 MITIGATING ALARM DISRUPTION

The effect of interruptions on task performance may be mediated by the information carried by the interrupting cue. Often, alerts indicate the state of the world in a binary fashion indicating that things are either safe or they are not. An alarm will sound if an automated aid diagnoses a situation to be dangerous, but otherwise the aid will remain silent, indicating an “all safe” state. While these aids have been demonstrated to be useful compared to no alarm, the binary nature of the typical alarm leaves the human with few options regarding when, or if, to respond. In contrast, **likelihood alarms**, or alarms that express some confidence in their diagnostic decision, may help diminish performance costs associated with ongoing and automated tasks (Woods, 1995; Sorkin, Kantowitz & Kantowitz, 1988). If the human knows how urgent the alarm is, a more educated decision can be made as to if, or when to attend to it (Sorkin & Woods, 1985). Automated diagnostic information can be conveyed in a variety of ways that may help optimize task management. Likelihood alarms can exploit both visual or auditory stimulus characteristics to convey more specific diagnostic information compared to binary alerts.

Woods (1995) argued that it is important for the human to know how urgent the event signaled by the alarm is relative to other ongoing tasks. This is important because if the alert is signaling an event that is less important than attending to the ongoing task, then switching attention to perform the interrupting task will be inappropriate and the ongoing task will be degraded for no reason. However, if the person can assess whether it will be worthwhile to disengage from the ongoing task to attend to the alarm before having to do so, performance on both tasks should be more optimal. Woods (1995) suggests that alarms that allow for such **preattentive referencing** allow the human more flexibility in managing ongoing and interrupting tasks. Woods (1995) proposes that one way of achieving this preattentive advantage in alarm design is to use a multi-level **likelihood** alarm that allows the system to express its own degree of certainty of the risk associated by the signaled event.

Only two studies have compared likelihood with binary alarms. Sorkin, Kantowitz and Kantowitz (1988) examined the issue of likelihood alerts in a dual-task situation. The authors hypothesized that the informativeness of a cue would interact with overall task demands such

that when the alarm priority was low, the operator would only check the monitored channel if he or she was not busy with the primary task, but higher level alerts would generate rapid attention shifts regardless of primary task work load. The authors also thought that likelihood alerts would be more beneficial at more variable levels of primary task load. In addition, any time the human could check the monitored domain, the likelihood information displayed there would essentially help them check their own diagnosis of the situation and thereby aid situation awareness (Sorkin et al., 1988).

Subjects in Sorkin et al.'s (1988) experiment performed a primary continuous tracking task that was either easy or hard and a secondary diagnostic decision task. The human was aided on the secondary task by an automated monitoring and alarm system which was presented either visually or with speech and was either binary or had 4 states representing different levels of diagnosis (silent/white, "possible signal"/green, "likely signal"/yellow, "urgent signal"/magenta). The workload of the primary task was also manipulated. The authors found no effect of alert modality or resolution (type) on tracking performance, but did find that when tracking difficulty was high, the likelihood alert supported better performance on the alerting task itself. Performance on the latter was not influenced by alert modality.

St. John and Manes (2002) found that likelihood information displayed with one of the automation system further facilitated search. The authors note however, that search strategies became more complex with the more informative likelihood automation compared to the binary automation. The results of this study are enlightening, but not fully informative for the present hypothesis because the authors did not measure concurrent task performance.

5.0 CONFLICT DETECTION

The Cockpit Display of Traffic Information helps pilots detect aircraft that pose a collision threat to their own aircraft. There have been several studies that have examined CDTI use related to conflict resolution (Alexander, Wickens, & Merwin, 2005; Scallen, Smith, & Hancock, 1996; Wickens, Gempfer, & Morpew, 2000; Wickens, Helleberg, & Xu, 2002). Fewer studies have examined conflict detection with CDTI (Merwin & Wickens, 1996; Xu, Wickens, & Rantanen, 2004), and there have been no studies that have examined the effects of the use of a CDTI with an embedded alarm in multiple task situations. Those studies that have focused on conflict detection indicate systematic pilot biases when using CDTI visual information (Xu, Wickens, & Rantanen, 2004). These biases were reduced, and conflict detection accuracy increased, when an automated likelihood alert was embedded in the CDTI. However, it is reasonable to assume that there may be performance consequences associated with the use of a CDTI alerting system for ongoing tasks.

The use of a CDTI alerting system to examine the effects of interruptions on ongoing task performance has several benefits. First, the CDTI alerting system is one that works in conjunction with a human pilot. The CDTI allows the human to access visual information about air traffic (the raw data) so the human can devote some attention to monitoring for collision threats. Second, the cost of missing an event within the context of an aircraft collision is high. That means the CDTI system allows for a meaningful manipulation of the threshold level of the embedded alert that is directly related to a real-world context, and consequently makes the results more generalizable to real-world situations. That is, the high cost of automation misses dictates a low threshold setting such that false-alerts will be frequent. In addition, when events to be detected by automation are rare, as is the case in airspace conflicts in aviation, automation false alarms rate is particularly higher in proportion to hits compared to automation monitoring for more frequent events (Parasuraman, Hancock, & Obofinbaba, 1997; Krois, 1999). Furthermore, it is the case that many issues of how to design such an alerting system remain to be resolved. These include not only the appropriate threshold setting, but also the modality and the resolution (likelihood vs. binary). Therefore the CDTI alerting system offers an excellent domain to evaluate how alert characteristics mediate the disruptive effects of interruptions due to high false alarm rates on ongoing task performance.

6. SUMMARY

6.1 Purpose

Several studies have compared auditory vs. visual interruption effects in a dual task context (e.g., Wickens & Liu, 1988; Latorella, 1996; Banbury et al., 2001; Ho, Nikolic & Sarter, 2003; Iani & Wickens, 2004) and have revealed a pattern we call “auditory preemption” in which auditory alerts tend to drive attention rapidly to the alerted domain, but at a cost to ongoing visual tasks. However none of these studies have examined this issue in the context of a specific alerting system, in which the threshold (ratio of FA to misses) is varied. Furthermore none of the studies which have manipulated the threshold of an alerting system have done so while correspondingly varying the interruptability of an ongoing task (i.e., working memory load or stability). And, those studies which have varied the alerting threshold in a dual-task context, where the full implications of reliance and compliance could be examined (Dixon & Wickens, in press; Wickens et al., 2005; See Wickens & Dixon for summary) have examined neither the likelihood alert nor the modality. Also, several studies have considered explicitly different characteristics of the ongoing task in an interruption study (Bailey et al., 2003; Gillie and Broadbent, 1989 Iani & Wickens, 2004), and while some have varied the modality of the interrupting task, only one has done so in the context in which the interrupting task was an alerting task (Iani & Wickens, 2004). And, in this study, the alert task was perfect, so miss - false alarm tradeoffs could not be examined.

Finally, only two studies appear to have compared a likelihood alert with binary, on-off alert (Kantowitz et al., 1988; St. John & Manes, 2002), and while modality was varied in the former, only a single ongoing task was employed, and the alert threshold remained constant, so effects within the compliance-reliance framework could not be examined. The latter study was in a single task paradigm. Thus, the general goal of the proposed research is to find out if the factors, of auditory alert preemption, ongoing task interruptability, compliance and reliance-mediated threshold setting, and likelihood alert benefits, all of which have been examined individually or sometimes in pairs, can “scale up” to account for performance variance in a somewhat realistic multi-task environment, which would be typical of the airplane cockpit where alerting tools are proposed to be of great value.

6.2 Hypotheses

In four experiments pilots performed a CDTI air traffic conflict monitoring task, assisted by an imperfect alert (the interrupting task or IT), in the context of two different forms of ongoing tasks (OT: tracking and working memory), that vary in their interruptability. The variables that we manipulate across all experiments (modality, concurrent task, alert threshold, and alert type) allow us to frame a set of hypotheses that can be expressed as a series of predicted main effects or interactions, whose influence is predicted to be observed on both the ongoing task (OT) and the interrupting task (IT):

H1: Ongoing Tracking Task Stability: We predict that an unstable tracking task will resist interruptions from the CDTI alert more than a stable tracking task because pilots will be less willing to leave the unstable tracking task. This prediction should result in slower response times to CDTI conflicts in the unstable tracking condition compared to the stable tracking condition.

H2a: Ongoing Working Memory Task: We predict that the ongoing auditory working memory task will resist interruption more than the visual working memory ongoing task because pilots will be able to revisit the visual working memory task and therefore the cost of leaving this task to attend to an alert will be less than with the auditory working memory task. If this is so, slower response times to CDTI conflicts should ensue in the auditory working memory condition compared to the visual working memory condition because of pilots' reluctance to switch in the former until the task is completed

H2b: Ongoing Auditory Working Memory Task: performance on an **ongoing auditory working memory task** will be affected by when during the task an alert is presented. We expect that working memory accuracy will be most negatively affected when an alert is presented early or in the middle of the working memory task because (a) pilots would be likely to switch immediately to the alert task, since to delay response to the alert for the entire working memory segment would risk missing a conflict, and (b) pilots will likely feel that memory load is still low enough to respond to the alert. We expect that factors that delay the switch (in order to stay with the working memory task) will thereby increase auditory working memory task performance.

H3: Interrupting Task Modality: We predict that auditory alerts will cause auditory preemption. Evidence of this should be in faster response times to alerts and possibly more accurate conflict detection with auditory alerts compared to visual alerts. In addition, concurrent task tracking error should be greater with auditory alerts compared to visual alerts because a rapid switch of attention to the automated domain will degrade concurrent task performance.

H4: Interrupting Task Threshold Setting: Our predictions regarding the impact of an ecological (lowered) alert threshold, resulting in an increased FA rate and decreased Miss rate, on both the IT and the OT are twofold:

- 1) Reduced **compliance** or “cry wolf” will increase response times and decrease sensitivity to CDTI conflicts.
- 2) Increased **reliance**, caused by fewer misses in the low threshold condition, will improve tracking performance.

H5a: Interrupting Task Alert Type: We hypothesize that likelihood alarms will support better performance than binary alarms. Specifically, we expect that likelihood alerts will enable better task management, so while response times to the CDTI alert may not be faster (or may even be a bit slower) in the likelihood condition, we expect that conflict detection accuracy should be preserved. Most importantly, we expect that likelihood alerts, by enabling the pilot to delay response to the alerted task in some cases, will allow the pilot to better protect ongoing task performance. Therefore, we expect reduced tracking and working memory error with likelihood alerts compared to binary alerts. These effects should be most manifest for those conflict trials where the likelihood alert makes its unique contribution: the "mid-level" alert.

H5b: Interrupting Task Alert Type: In addition, we hypothesize that likelihood alarms representing good human factors design, will mitigate other problems associated with interruptions and task difficulty. Specifically, we expect that costs associated with auditory preemption, unstable tracking, and working memory load, should be mitigated by the likelihood alert. This hypothesis predicts the particular form of interaction between alert type and a second variable such that the cost due to a second variable is less, or even beneficial, with the likelihood alarm relative to the binary alarm.

H6: Multiple Resource Effects: Consistent with Multiple Resource Theory (Wickens, 2002), we expect an advantage for auditory alerts with visual concurrent tasks and an advantage for visual alerts with auditory concurrent tasks. Multiple resource theory makes similar predictions to auditory preemption (H3) with regard to the interrupting task, but opposite predictions with regard to the OT modality.

In order to examine these issues, four experiments were conducted. In all four experiments, pilots from the University of Illinois performed two tasks simultaneously. One of the tasks was always a conflict detection task supported by a CDTI with a discrete imperfect alert. The second task was either a compensatory tracking task (Experiments 1 & 2) or an auditory (Experiment 3) or visual (Experiment 4) computational task. When displayed auditorally, the computational task has substantial working memory components. The tracking task had either a stable (interruptible) or unstable, (interruption resistant) version in Experiments 1 and 2 and these experiments were identical except that a neutral alert threshold (equal automation misses and false alarms) was used in Experiment 1 and a more ecological (lowered) alert threshold was used in Experiment 2 that minimized automation misses, but generated more automation false-alerts. Experiments 3 and 4 were identical to Experiment 2 except that pilots completed a navigation/communications computational task (auditory or visual respectively) instead of a compensatory tracking task. Alerts were delivered at three different times during each working memory problem set.

7. METHODS: EXPERIMENT 1

7.1 Participants

Twelve pilots from the University of Illinois Institute of Aviation were recruited to participate in the experiment. Participants ranged in age from 19 to 26 with a mean age of 22.6. Participants had normal, or corrected to normal vision. Participants had an average of 97 hours of flight experience. Participants were paid \$9 per hour for participating. Total participation time did not exceed 1.5 hours.

7.2 Design

The purpose of Experiment 1 was to examine the impact of interruptions, introduced by an alarm, on ongoing task performance. In addition, we were interested in determining if a likelihood alarm would mitigate the impact of imperfect alerting. To do this, we manipulated how interruptible (difficult) the ongoing task was, the type of alarm (likelihood vs. binary) and the alert modality (auditory vs. visual).

Table 2 provides a complete list of all conditions in Experiment 1. A two (interruptible or non-interruptible ongoing tracking task) x two (visual vs. auditory alert) x two (binary or likelihood alarm) within subjects design was used. Participants completed two sessions. The sessions were counterbalanced such that half of the participants were presented with a binary alarm in the first session and the likelihood alarm in the second session and this order was reversed for the other half of participants. The order of tracking difficulty and alert modality were also counterbalanced across participants.

Table 2: List of the conditions for Experiment 1.

Session 1 : Binary Alert
Interruptible Stable tracking/ auditory alert
Interruptible Stable tracking/visual alert
Non-interruptible Unstable tracking/visual alert
Non-interruptible Unstable tracking/auditory alert
Session 2: Likelihood Alert
Interruptible Stable tracking/ auditory alert
Interruptible Stable tracking/visual alert
Non-interruptible Unstable tracking/visual alert
Non-interruptible Unstable tracking/auditory alert

7.3 Procedure

Ongoing Task

Figure 3 illustrates the display that was used in Experiment 1. Participants performed an ongoing, first order, compensatory tracking task with a bandwidth of 0.30 hz that was presented centrally on the computer screen. The tracking task required participants to keep a cursor within an acceptable position inside a target rectangle presented on the computer screen. Participants controlled the position of the cursor with a joystick using their left hand. The interruptability of the tracking task was manipulated by making the task self-correcting or not. In the self-correcting condition, when the task was neglected, the error reached a peak value, but eventually the cursor moved back to its central position. In the non self-correcting, unstable condition, incorporating a positive feedback loop, the deviation of the tracking task error continued to grow unless corrected by the subject.

Interrupting Task: CDTI Conflict Alarm

As shown in Figure 3, a simple CDTI display was utilized for the interrupting task and was presented in the upper right hand corner of the screen. The CDTI monitored for potential collision threats and warned participants with a visual or an auditory alert if a collision threat was predicted. When conflicts were detected by the participant, participants were instructed to click on the intruding aircraft’s icon with the left or right mouse button to indicate which direction the aircraft should be routed to in order to avoid a conflict with their own aircraft. (While this

particular action is non-ecological – the pilot would not request maneuver of an intruding ship – it was designed to force the pilot to attend to the geometry of the conflict).

Figure 3: Rendering of the display presented to participants in Experiments 1 & 2



Each group of subjects participated in 4 separate conditions per session and each condition lasted for approximately 14 minutes. Figure 4 and Figure 5 illustrate the timing of the CDTI events for each condition. During each trial, a new aircraft appeared on the screen every 10 seconds in a continuous stream, for a total of 80 events per condition. There were never more than 4 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, right, and passing in front and behind ownship. Of the events, 50 % were conflict events. The conflict events were manipulated to represent a range of threat seriousness according to closest point of approach. Pilots were to judge a “conflict” to be any aircraft that would penetrate their protected airspace (3 nautical miles). As seen in Figure 3, a standard 3 mile ring was placed around ownship to support this judgment. If an alert was presented, it occurred 6 nautical miles from ownship at a constant 5 seconds before the point of closest approach was reached.

Figure 4: This figure shows the potential timing of conflict and non-conflict events and alert onsets. Plane (P) arrivals are numbered sequentially.

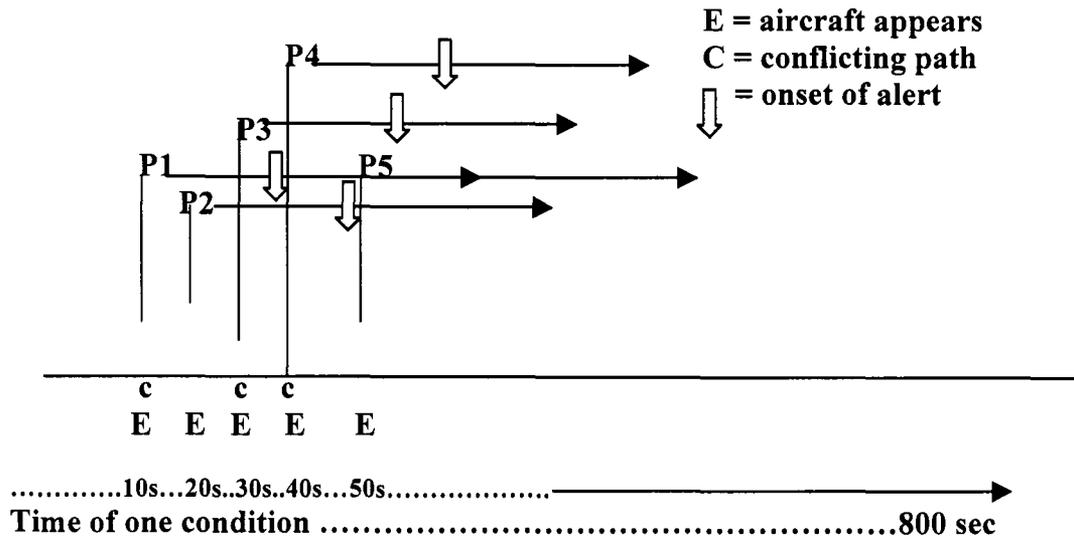
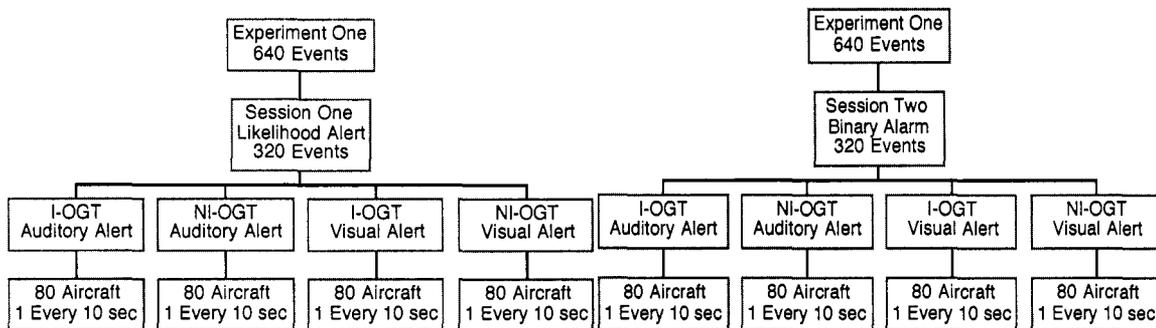


Figure 5: This figure illustrates the characteristics of the CDTI alarm as a function of experimental condition.



Tables 3 and 4 illustrate the number of automation decisions (hits, misses, correct rejections, and false alarms) for the binary and likelihood alarm respectively. The reliability of the CDTI alarm was less than perfect, which meant that in addition to sampling the display to detect visual alerts and listening for auditory alerts, participants needed to sample the raw data on the display to pick up possible conflicts the automation missed and to check the authenticity of the CDTI alerts that could be false and to possibly detect an emerging conflict before a true alarm was presented. In Experiment 1 the alert showed a slight bias for alarms, but had a miss to false-alarm ratio of 1:1.

Table 3: The number of Hits, False-Alarms, Misses, and Correct Rejections for the Binary Alarm in Experiment 1.

Experiment 1	Threat	No Threat
Alarm	36 H	10 FA
No Alarm	10 M	24 CR

Table 4: The number of Hits, False-Alarms, Misses, and Correct Rejections for the Likelihood Alarm in Experiment 1.

Alert Says	Danger	Modest	Safe
Danger	27	3	4
Modest	3	15	3
Safe	4	3	18

As illustrated in Table 5, the alarm was further manipulated to include auditory and visual alarm condition. The characteristics of the alarm system were taken from Xu, Wickens, and Rantanen (2004). The visual alert for the 2-level (binary) alarm indicated a threat by a red square onset around the entire CDTI display as well as a change in color of the intruding aircraft icon from green to red. The auditory alert for the binary alarm indicated a threat by the presentation of a synthesized voice that said “conflict conflict”.

Table 5: Likelihood alarm characteristics associated with each level of threat.

Three-level MD Alert: Taken directly from Xu, Wickens, & Rantanen (2004)

Alert level	Color of traffic icon	Auditory Alert	MD (mile)
No alert	Green	silent	> 3
Medium alert	Orange	“traffic traffic”	3-4
High alert	Red	“conflict conflict”	< 3

The visual alert for the two-level likelihood alarm indicated a threat by a red square color onset around the CDTI display for high level conflict predictions and an orange square color onset around the CDTI display for a mid-level conflict predictions. Both of these visual alerts were accompanied by a change in color of the intruding aircraft from green to red or green to orange depending on the severity of the potential conflict (in terms of closest point of approach). The auditory alert for the likelihood alarm was the presentation of a synthesized voice saying either “conflict conflict” for a high alert, or “traffic traffic” for a less severe alert (again, based on closest point of approach to ownship). The auditory alerts were also accompanied by a change

in the intruding aircraft icon from green to orange (mid-level alert) or red (high-level alert). The diagnoses of no threat, medium threat, and high threat were determined by the projected closest distance an aircraft would come to intruding on ownships' airspace (Xu, Rantanen & Wickens, 2004). The medium alert category was equally divided into 3 and 4 mile separations at the point of closest approach. Identical sequences of conflict geometry were presented in the likelihood and binary condition, so the only difference between these was that, for the likelihood alerts, all of these mid level (3 or 4 mile) conflicts triggered an amber (or "traffic") alert, whereas for the binary condition the 3 mile conflicts only triggered the alert (red or "conflict"). Thus, in the likelihood condition pilots should systematically indicated a "conflict" (by clicking on the airplane icon) on half of the mid-level alerts. Those with a 3 mile projected separation. Four mile and greater separation required no response.

7.4 Pilot Study

Prior to the four main experiments, a pilot study was conducted to assure equal salience of the visual (color) and auditory alerts. Six participants focused attention on the middle of the tracking display and made simple and choice RT responses with a key press to either the (572ms)amber stimulus pair, or the "traffic" "conflict" stimulus pair. Results of this study revealed significantly **shorter** response times with the visual (489ms) than the auditory (572ms) events ($t(1, 5) = 4.34, p < .05$), thereby assuring that any subsequent costs that might be observed to the visual alerts, could not be the result of reduced salience due to their eccentricity in peripheral vision.

7.5 Dependent Measures

Dependent measures for the alerted domain included the time it took participants to detect conflicts. This response time was measured from the time an aircraft icon appeared on the screen until the participant clicked on it, indicating it was a threat. Conflict detection accuracy was also measured (Sensitivity: d') for each subject in each condition as was the percentage of time the participants' eyes dwelled on the tracking task compared to the CDTI task. For the ongoing (tracking) task, we measured each subjects tracking error for each condition. Eye movements were also recorded for ten of the participants.

8. RESULTS: EXPERIMENT 1

At the outset of all of our analyses, data were checked for normality. None of our data was significantly skewed and therefore did not require any transformation. Outliers (performance points greater or less than 2 SD) were deleted.

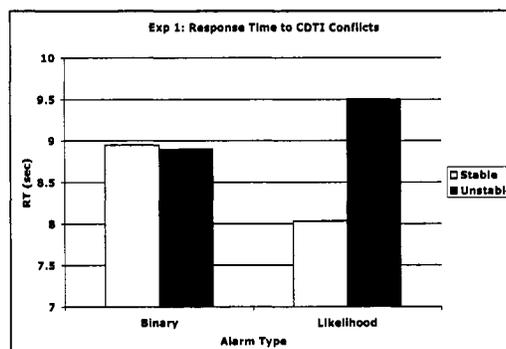
In Experiment 1 we manipulated two characteristics of a CDTI alarm system with a neutral threshold: the type of alarm (binary vs. likelihood) and the modality of the alert (auditory vs. visual). In addition, we manipulated the difficulty of the concurrent tracking task (stable vs. unstable dynamics).

8.1 Alerted Task Performance

Response Time

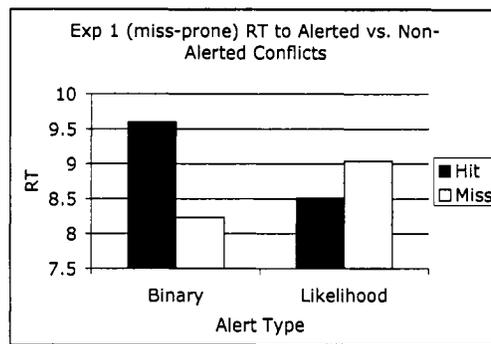
Participants had slower response times to CDTI conflicts during unstable tracking ($M = 9.22$) compared to the stable tracking ($F(1,11) = 7.02, p. < 0.05$). Participants were equally fast to respond to conflicts whether the alarm was presented auditorally or visually (color onset) ($F(1,11) = 0.45$), and whether the alarm was likelihood or binary. As seen in Figure 6, ongoing task (tracking) difficulty did interact with alarm type to impact reaction time ($F(1,11) = 8.62, p. < 0.05$) which indicated a facilitative effect of the likelihood alert with easy tracking, but a cost with the likelihood alert during difficult tracking. Alternatively, increasing tracking difficulty prolonged CDTI response times with the likelihood alert, but not with the binary alert.

Figure 6: Response time to CDTI conflicts for alarm type and tracking difficulty.



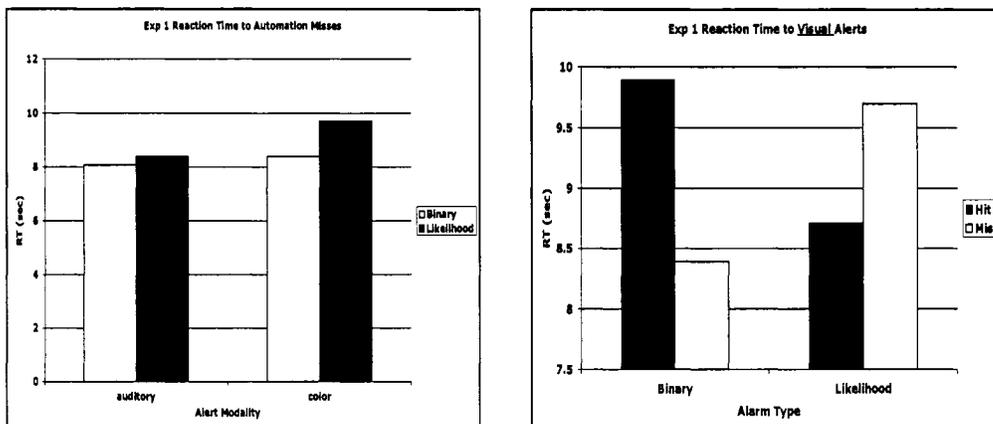
As shown in Figure 7, when responses were parsed by whether the participant was responding to an automation hit vs. an automation miss, a significant interaction emerged between alert type (binary vs. likelihood) and automation decision (hit vs. miss) such that participants were slower to respond to automation hits compared to misses in the binary alert condition but the reverse was true in the likelihood condition ($F(1,11) = 10.39, p. < 0.01$). There were no other significant two-way interactions.

Figure 7: This figure shows response times to CDTI conflicts as a function of alarm type and automation decision.



As shown in Figure 8, a three way interaction between alarm type, alert modality, and automation decision was significant ($F(1,11) = 4.86, p. < 0.05$) and suggests that the tradeoff interaction depicted in Figure 7 is most prominent in the visual alerts (right panel of Figure 8). There were no other significant three-way interactions.

Figure 8: The panel on the right shows RT to conflicts in the visual alert condition as a function of automation decision and alarm type. The panel on the left shows RT in the auditory alert condition.



Although the benefits associated with the likelihood alert seem ambiguous, we predicted that the likelihood alert would aid task performance by allowing participants to prioritize tasks more optimally. More specifically, we thought that the mid-level likelihood alert should lead to prolonged reaction times to the CDTI task because participants should feel less urgency to switch to the alerted task. This should, in turn, preserve ongoing (tracking) task performance. To examine this issue, we compared response time for high level and mid-level likelihood alerts. Participants were indeed slower to respond to conflicts signaled by a likelihood mid-level alert ($M = 10.15s$) compared to those signaled by a likelihood high level alert ($M = 7.45$) ($F(1,12) = 42.58, p < .05$). We also compared the mid-level likelihood alert to the same conflicts signaled by the binary alert. Here we found no statistical difference in response time between the mid-level likelihood alert ($M = 10.15$) and the same events signaled by the binary alert ($M = 9.45$), although the trend was in the direction we would expect, with longer response times to the mid-level likelihood alert ($F(1,11) = 2.23, p > .05$).

Sensitivity

Increased tracking difficulty had no impact on participants' ability to detect CDTI conflicts, $F(1,11) = 2.20, p > 0.10$. Participants were more accurate in detecting CDTI conflicts in the auditory condition compared to the visual alert condition ($M = 2.14$ vs. $M = 1.37$; $F(1,11) = 54.07, p < 0.01$). Alarm type (binary vs. likelihood) did not impact participants conflict detection accuracy, $F(1,11) = 0.14, p > 0.10$. There were no other significant main effects or interactions.

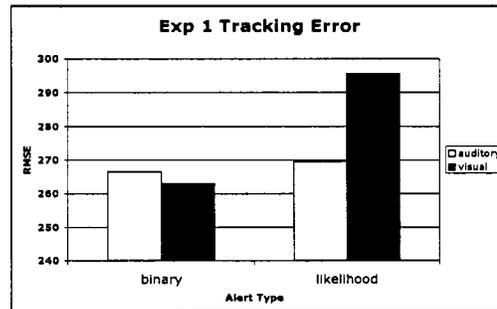
8.2 Concurrent task Performance

Tracking Error

When the difficulty of the tracking task was increased, participants' tracking error almost doubled ($M = 193.47$ vs. $M = 353.93$) ($F(1,11) = 372.78, p < 0.01$). Tracking performance was the same independent of the presentation modality of the alert. However, tracking performance was worse during the likelihood alerting condition compared to the binary alerting condition ($M = 282.25$ vs. $M = 264.5$) ($F(1,11) = 8.26, p < 0.05$).

Figure 9 below shows an interaction between alert type and modality ($F(1,11) = 6.13, p < 0.05$) such that concurrent tracking was particularly hurt by the likelihood alert when it was delivered visually. There were no other significant main effects or interactions.

Figure 9: Tracking error (RMSE) as a function of alarm type and alert modality.



We found no statistically significant difference in tracking error when we compared mid-level and high level alerted trials in the likelihood condition. We also found no difference in tracking error during mid-level likelihood alerts compared to the same events signaled by the binary alert.

9. METHODS: EXPERIMENT 2

9.1 Participants

Twelve pilots from the University of Illinois Institute of Aviation were recruited to participate in the experiment. Participants ranged in age from 20 to 26 with a mean age of 23.1. Participants had normal, or corrected to normal vision. Participants had an average of 102 hours of flight experience. Participants were paid \$9 per hour. Total participation time did not exceed 1.5 hours.

9.2 Design

Experiment 2 was identical to Experiment 1 except that the detection threshold of the aid was changed to induce a more ecologically valid ratio of automation errors. The aid for Experiment 2 reflected a bias towards false alarm errors. This allowed us to determine the impact of interruptions due to the high false alarm rate but lower miss rate characteristic of low alert thresholds often chosen for alert systems that detect rare, but potentially catastrophic events, such as the CDTI.

The alarm threshold for Experiment 2 resulted in a ratio of false alarms to misses of 4:1 while the reliability of the system remained essentially the same as it was in Experiment 1. The categorization of collisions and automated decisions is outlined in Table 6 for both the binary and likelihood alerts.

Table 6: This table lists the CDTI alert decisions for the Binary and Likelihood Alerts.

Binary Alarm

Experiment 2	Danger	Safe
Alarm	36 H	16 FA
No Alarm	4 M	24 CR

Likelihood Alarm

Alert Says	Danger	Modest	Safe
Danger	27	4	8
Modest	1	15	4
Safe	2	1	18

10. RESULTS: EXPERIMENT 2

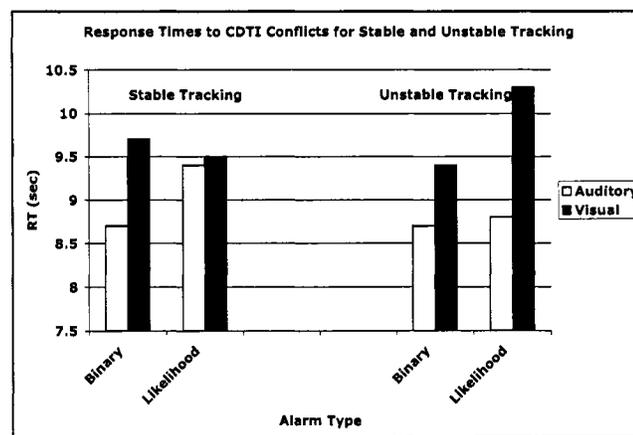
Experiment 2 was identical to Experiment 1 except that the alert threshold was lowered to simulate real world systems that strive to minimize misses. The false alarm to miss ratio in Experiment 2 was 4:1.

10.1 Alerted Task Performance

Response Time

Increased tracking difficulty did not directly affect the time it took participants to respond to CDTI conflicts ($F(1,11) = 0.05$). There was a marginally significant effect of alarm modality with faster responses to auditory alerts ($M = 8.9$) compared to visual alerts ($M = 9.5$) $F(1,11) = 4.09$, $p = .068$. There was no main effect of alarm type (binary vs. likelihood) on response times to the conflicts ($F(1,11) = 0.16$). There were no significant two-way interactions between alarm type, alert modality or difficulty when all responses were grouped. However, there was a significant three-way interaction between alarm type, modality and tracking difficulty ($F(1,11) = 5.2$, $p < .05$). As shown in Figure 10, as tracking became more unstable, a cost for the visual likelihood alarm emerged that wasn't present with the stable tracking task.

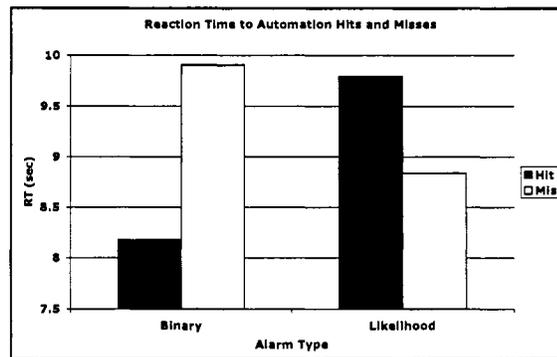
Figure 10 . Three way interaction between alarm type, teaching difficulty and modality.



As shown in Figure 11, when responses were parsed by whether participants responded to an automation hit vs. an automation miss, a significant interaction emerged between alarm type

(binary vs. likelihood) and automation decision ($F(1,11) = 11.07, p < 0.01$) such that in the binary alarm condition, response times to automation hits were faster than response times to automation misses but this trend was reversed in the likelihood condition. We note the reversal of the form of the interaction, from that shown for Experiment 1, in Figure R2. There were no other significant interactions.

Figure 11: Response times to CDTI conflicts as a function of Alarm type and automation decision.



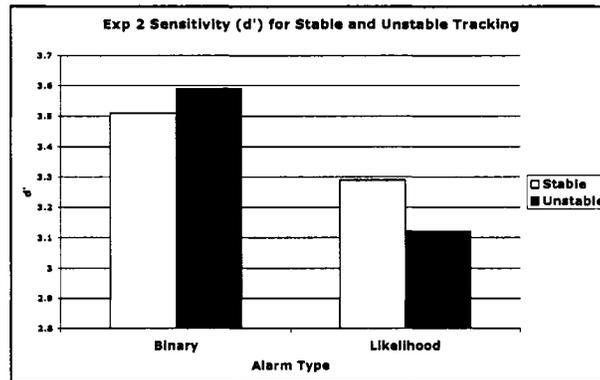
When we compared response times to the mid-level likelihood alert to those of the high level likelihood alert (as we did in Experiment 1) we found that participants were significantly slower to respond to mid-level alerts compared to high level alerts ($M = 13.15s$ vs. $M = 9.58s$) ($F(1, 12) = 37.83, p < .05$), replicating our finding of Experiment 1. In addition, we compared response times to potential conflicts signaled by the mid-level likelihood alert with the same events signaled by the binary alert. Note, these aircraft had the same trajectory and reached the same closest point to ownship: the only difference was the type of alert that signaled the potential conflict. Here we found that participants took longer to respond to conflicts signaled by the likelihood mid-level alert ($M = 12.91s$) compared to the same conflicts signaled by the binary alert ($M = 9.80$) ($F = 35.31, p < .05$).

Sensitivity

Increased tracking difficulty had no direct impact on participants' sensitivity ($F(1,11) = 0.24, p > 0.10$). However, as depicted in Figure 12, tracking difficulty did interact with alarm

type ($F(1,11) = 4.69$, $p. = 0.05$) such that difficult tracking was especially damaging to sensitivity in the likelihood condition, and that the likelihood cost was greater with difficult tracking.

Figure 12: Sensitivity (d') as a function of alarm type and tracking difficulty.



Participants were marginally more sensitive to auditory alerts compared to color alerts ($M = 3.41$ vs. $M = 3.34$: $F(1,11) = 4.57$, $p. = 0.056$). There were no other significant main effects or interactions.

10.2 Concurrent task Performance

Tracking Error

Increased tracking difficulty increased tracking error ($M = 172.45$ to $M = 316.50$: $F(1,11) = 675.66$, $p. < 0.01$). Also, there was a marginally significant cost to tracking performance when and auditory alert was presented compared to when a visual alert was presented ($M = 250.36$ vs. $M = 238.60$: $F(1,11) = 3.59$, $p. = 0.09$). The type of alarm (binary vs. likelihood) had no impact on participants' tracking error ($F(1,11) = 0.08$, $p. > 0.10$). There were no significant interactions between alert modality, alarm type or tracking difficulty on tracking performance.

As in Experiment 1, we found no statistically significant difference in tracking error when we compared mid-level and high level alerted trials in the likelihood condition. We also found no difference in tracking error during mid-level likelihood alerts compared to the same events signaled by the binary alert.

11. RESULTS: EXPERIMENT 1 VS. EXPERIMENT 2

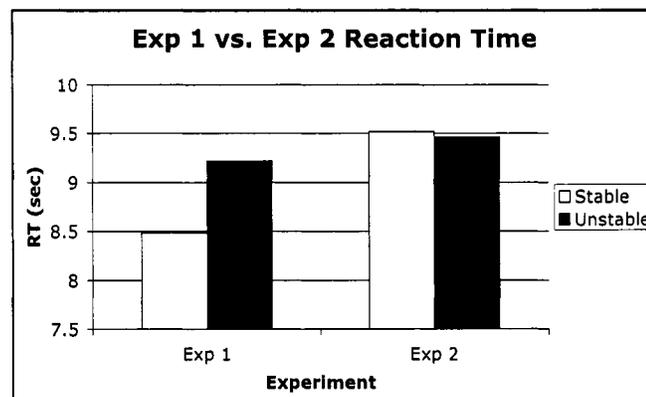
In Experiment 2 the threshold of the alert was lowered to simulate real-world alarm systems which typically chose to minimize misses (with the consequence of making the system false-alarm prone). The same analyses that were run for Experiment 1 and Experiment 2 independently, were run again using experiment as a between subjects factor in order to determine performance differences that arose due to the lowered threshold of the alarm (decreased miss rate, increased false alarm rate), therefore examining hypothesis 4.

11.1 Alerted Task Performance

Response Time

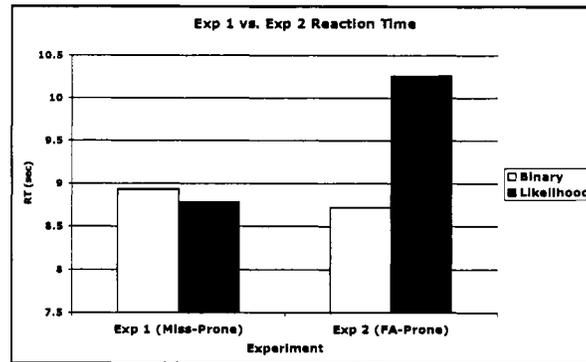
Lowering the alert threshold had no overall impact on participants' response time to CDTI conflicts, ($M = 9.49$ vs. $M = 8.86$). $F(1,11) = 0.22$, $p. > 0.10$. However, Figure 13 shows that tracking difficulty did interact with alert threshold (neutral in Experiment 1 and low Experiment 2) ($F(1,11) = 4.37$, $p. < 0.05$) indicating that the threshold reduction increased RT only when the tracking was easy.

Figure 13: Response times to CDTI conflicts as a function of alert threshold (experiment) and tracking difficulty.



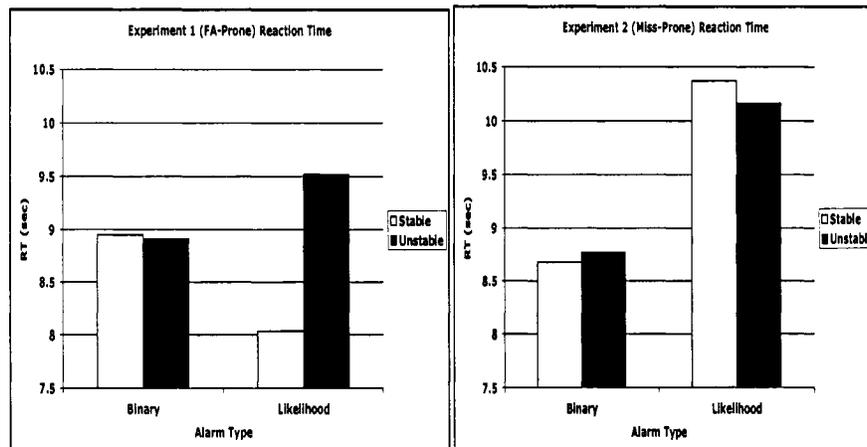
As shown in Figure 14, there was also a marginally significant interaction of experiment with alert type (binary vs. likelihood ($F = (1,11) = 3.42$, $p. = 0.07$)), indicating that the false-alarm increasing threshold shift only slowed RT when the alert was a likelihood alert.

Figure 14: Response times for CDTI conflicts for Experiment 1 & 2 as a function of alarm type.



Finally, as shown in Figure R15, there was a three-way interaction between tracking difficulty, alert threshold, and alarm type (binary vs. likelihood), $F(1,11) = 5.76$ $p. < .05$, such that a likelihood alert cost emerged in Experiment 2, especially with the stable tracking task. Alert threshold (neutral vs. low/false-alarm prone) did not significantly interact with alert modality ($F(1,11) = .51$).

Figure 15: Response times for CDTI conflicts for Experiment 1 (left panel) and Experiment 2 (right panel) as a function of alarm type and tracking difficulty.

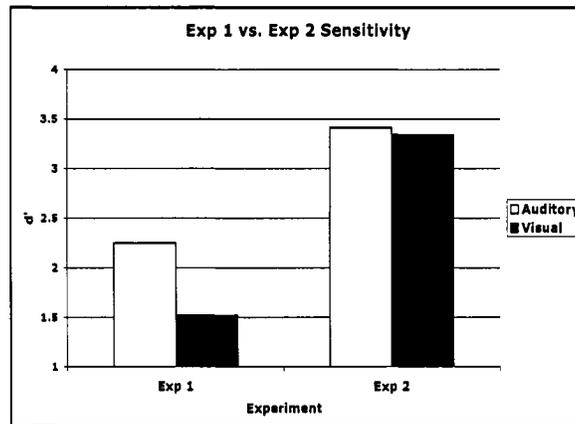


Sensitivity

Participants were much more accurate in detecting CDTI conflicts in Experiment 2 compared to Experiment 1 ($M = 3.38$ vs. $M = 1.89$; $F(1,11) = 65.71$, $p. < 0.01$) even though the

alerting systems themselves were nearly equally sensitive (1.33 vs. 1.54 for Experiments 1 & 2 respectively). While tracking difficulty did not interact with alert threshold (experiment) to impact participants accuracy in detecting CDTI conflicts, $F(1,11) = 0.073$, $p. > 0.05$), Figure 16 depicts a significant interaction between modality and alert threshold ($F(1,11) = 39.71$, $p. < 0.01$) such that participants accuracy in detecting CDTI conflicts was helped more by the lowered threshold shift in the visual alert condition compared to the auditory condition. Conflict detection accuracy was essentially equivalent across modality in Experiment 2. There were not other significant interactions across experiments (alert threshold).

Figure 16: Sensitivity (d') for Experiments 1 & 2 as a function of alert modality.

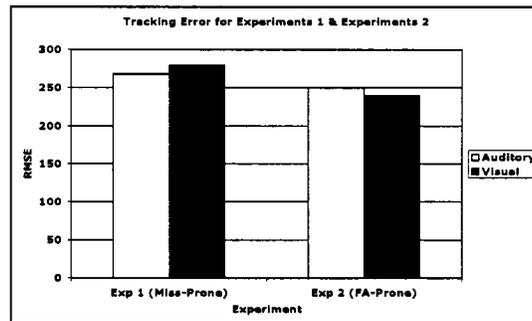


11.2 Concurrent Task Performance

Tracking Error

Tracking error was better in Experiment 2 compared to Experiment 1 ($M = 244.70$ vs. $M = 273.70$, $F(1,11) = 3.24$, $p. = 0.05$). Tracking difficulty did not interact with alert threshold to impact tracking error, $F(1,11) = 2.58$, $p. > 0.10$, nor did alarm type interact with alert threshold, $F(1,11) = 2.24$, $p.>.10$. But as shown in Figure R17, between experiment analyses did reveal an interaction between modality and alert threshold (experiment) ($F(1,11) = 4.64$, $p. < 0.05$) such that the decrease in tracking error with the threshold shift from Experiment 1 to Experiment 2 was greater with visual alerts than with auditory alerts.

Figure R17: Tracking Error for Experiments 1 & 2 as a function of alert modality.

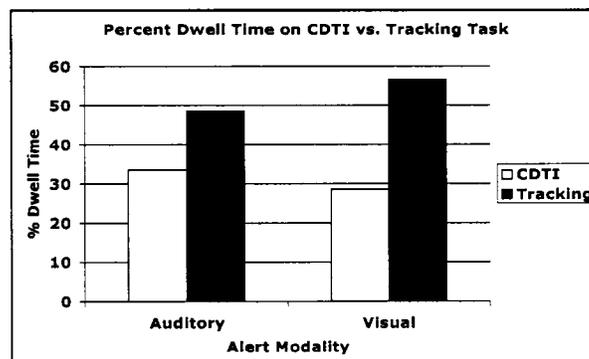


Eye Scanning

Visual scanning was analyzed as the percentage dwell time on either the tracking or CDTI area of interest, as a function of experiment. There was an expected effect that visual attention was more occupied with the tracking task (53%), imposing continuous visual demands, compared to the CDTI task (28%), whose demands were intermittent ($F_{20}=16.37, p<.01$) However, there was no effect of experiment. Decreasing misses did NOT increase reliance which would have made more attention available for the tracking task, an effect on scanning which, had it been observed, would have been consistent with the better tracking performance seen in Experiment 2.

As seen in Figure 18, there was also an interaction between modality and Area of interest such that the CDTI (white bar) demanded more attention when it was paired with an auditory alert, and the tracking display (blue bar) correspondingly received less attention ($F_{20} = 16.0, p<.01$) This is trend is consistent with the poorer tracking performance seen during auditory alerts.

Figure18: Percentage dwell time on the CDTI as a function of alert modality



While this finding of an auditory increase in CDTI attention demand may initially seem counter-intuitive, when one considers that the greatest source of visual information was always the visual data on the CDTI in both conditions (not the periodic automated alert) it is reasonable to consider that it was easier for subjects to integrate the visual alert information with the visual CDTI than the auditory alert information with the visual CDTI, so the latter required that more attention be directed to the raw data.

12. METHODS: EXPERIMENT 3

12.1 Participants

Twelve pilots from the University of Illinois Institute of Aviation were recruited to participate in the experiment. Participants ranged in age from 20 to 26 with a mean age of 23.5. Participants had normal, or corrected to normal vision. Participants had an average of 105 hours of flight experience. Participants were paid \$9 per hour for participating. Total participation time did not exceed 1.5 hours.

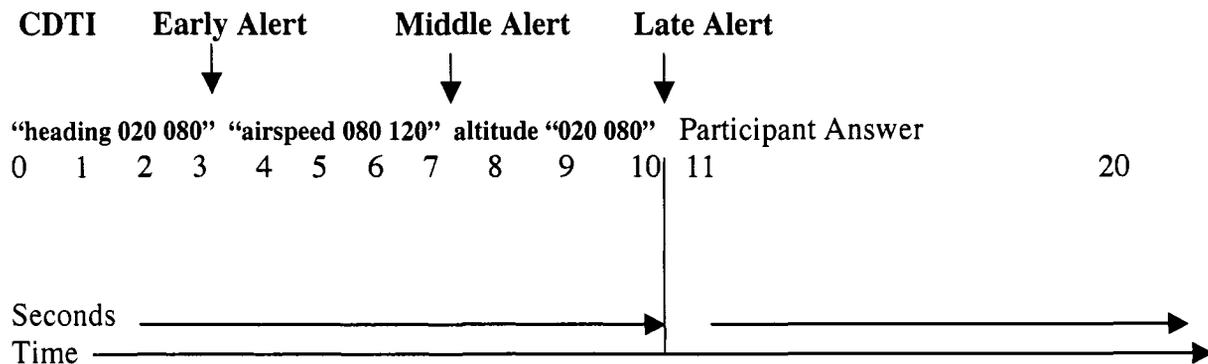
12.2 Procedure

Experiment 3 was identical in terms of the (CDTI) alerted domain to Experiment 2 (with the ecological alert threshold). The number and timing of air traffic events was also equivalent to Experiment 2, as were the alarm (binary vs. likelihood) and alert (auditory vs. visual) conditions. However, unlike Experiment 2, in which we used a compensatory tracking task, Experiment 3 employed an auditory computational working memory task as the ongoing task. We used the working memory task in order to assess the impact of imperfect alerting, and the possible mitigating effects of likelihood alarms, in a different task environment related to aviation than was used in the first two experiments. Whereas the concurrent tracking task of Experiments 1 and 2 were analogous to flight control, the concurrent auditory working memory task in Experiment 3 mimicked many of the demands of ATC communication of navigational information.

Another objective of Experiment 3 was to determine how different working memory demands affect both ongoing task performance when the task is interrupted by an alarm, and how these demands affect responses to the alerted domain. To examine these issues, we used a working memory task that progressively increased memory load interspersed by a low load interval. This structure should allow for short “windows of opportunity” during which time interruptions will be less disruptive. Figure 6 illustrates the timing of the working memory problems and the interjected CDTI alerts. The working memory task included a ten second presentation of three pairs of numbers related to current and desired heading, altitude, and airspeed. The participant was to listen to each pair of numbers, determine the absolute difference

between the numbers, and then vocally indicate their answer to the experimenter for all three sets during a ten second period of silence. For example, the participant would hear “heading, 270 (two, seven, zero), 200 (two, zero zero), altitude, 020 (zero, two zero), 080 (zero, eight, zero), airspeed, 130 (one, three, zero), 150 (one, five, zero). Only after all three problem sets were presented were the pilots to indicate their answer for all three. In this example the participant would say “70, 60, 20”, during the 10 second answer phase. If the participants forgot one or more of the problems, they were supposed to say “no value” in its’ place. The presentation of each number set took two seconds, with a two second delay between each problem, for a total of ten seconds. After each set of three problems (heading, altitude, airspeed), there was a ten second period of silence during which the participants were supposed to vocalize their answer. As shown in Figure 19, the CDTI alerts were timed to interject either early, in the middle, or late in the working memory task.

Figure 19: This figure shows the timing of the working memory task and the interjected CDTI alerts.



13. RESULTS: EXPERIMENT 3

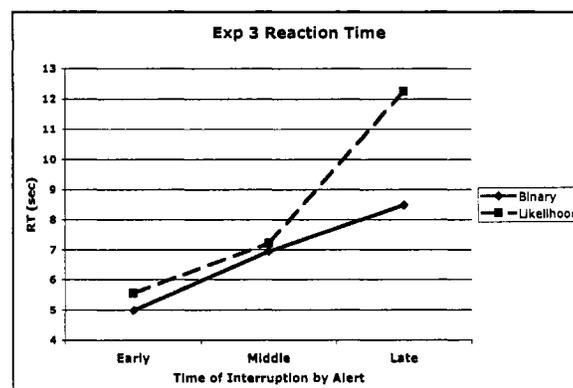
13.1 Alerted Task Performance

Response Time

Participants were faster to respond to CDTI conflicts signaled by a visual alert compared to those signaled by an auditory alert ($M = 7.74$ vs. $M = 8.13$: $F(1,11) = 5.51$, $p. < 0.05$), and faster to respond to CDTI conflicts in the binary alert condition compared to the likelihood alert condition ($M = 7.41$ vs. $M = 8.47$: $F(1,11) = 10.22$, $p. < 0.05$). Alert type (binary vs. likelihood) and alert modality (auditory vs. color) did not interact to impact reaction time to CDTI conflicts, $F(1,11) = 0.11$, $p. > 0.10$).

Response time data were then parsed based on when during the ongoing working memory task an alert was presented (towards the beginning = early, in the middle = middle, or towards the end = late). These data are shown in Figure 20. Response times increased as the alerts occurred progressively later in the concurrent task sequence, $F(1,11) = 29.80$, $p. < 0.01$. There was also a significant interaction between when during the working memory task an alert occurred (early, middle, late) and alarm type ($F(1,11) = 8.23$, $p. < 0.01$) such that participants were especially slow to respond to alerts presented late in the likelihood alert condition.

Figure 20: Response times to CDTI alerts at each interruption time as a function of alarm type.



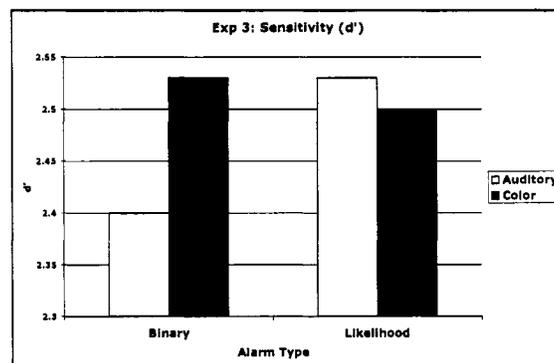
Response times to the mid-level likelihood alert were compared with responses from both the high-level likelihood alert and responses to events that were signaled by the binary alert but

were identical to those signaled by the mid-level alert in the likelihood condition. Consistent with Experiments 1 and 2, participants were slower to respond to the mid-level alert ($M = 10.85$) compared to the high level likelihood alert ($M = 8.04$) ($F(1,11) = 20.74, p < .05$). In addition, participants were slower to respond to conflicts signaled by the mid-level likelihood alert ($M = 10.85$) compared to identical conflicts signaled by the binary alert ($M = 7.4$) ($F(1,11) = 13.91, p < .05$), suggesting that it was the particular alert type and not the difficult to resolve conflict geometry that was responsible for the slowing.

Sensitivity

There was no main effect of either alert modality (auditory vs. color) ($F(1,11) = 0.18, p > 0.10$) or alarm type (binary vs. likelihood) ($F(1,11) = 0.13, p > 0.10$) on participants' accuracy in detecting CDTI conflicts. However, as shown in Figure 21 there was an interaction between alarm type (binary vs. likelihood) and modality (auditory vs. visual), $F(1,11) = 15.92, p < 0.01$. The interaction revealed auditory alerts, delivered in binary fashion, were particularly harmful to participants' accuracy. There were no other significant interactions.

Figure 21: Sensitivity (d') for Binary and Likelihood alarms as a function of alert modality.



13.2 Concurrent Task Performance

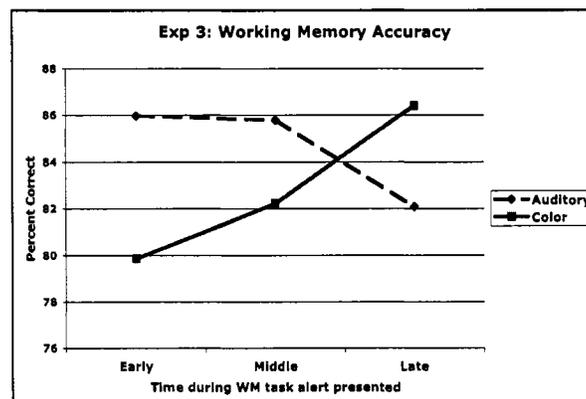
Working Memory Accuracy

Color CDTI alerts were slightly more disruptive than auditory alerts to participants concurrent working memory (communications) task performance (accuracy) ($M = 83.6\%$: vs. M

= 85.3%: ($F(1,11) = 6.35, p. < 0.05$). However, this effect of modality can only be interpreted within the context of the significant modality x alert type interaction ($F(1,11) = 9.73, p < 0.01$), which revealed a visual cost with the likelihood alert, but a slightly smaller visual benefit (auditory cost) with the binary alert.. There was no main effect of alert type (binary vs. likelihood) on concurrent working memory performance. ($F(1,11) = 0$).

As shown in Figure 22, when working memory performance data were parsed based on when during the task an alert occurred (early, middle, or late), auditory alerts did not degrade concurrent communication (working memory) task performance unless they were presented late in the task but the converse was true of color alerts $F(1,11) = 5.28, p. < 0.05$. That is, the cost to concurrent computation task performance was progressively reduced when color alerts were presented later in the task, Thus in the visual (color) alert condition, there was a reciprocity between performance on the CDTI (RT) task (Figure 20) and the working memory task (Figure 21), as the timing of the former is varied. In contrast, in the auditory condition, both tasks suffered as the alert was presented later.

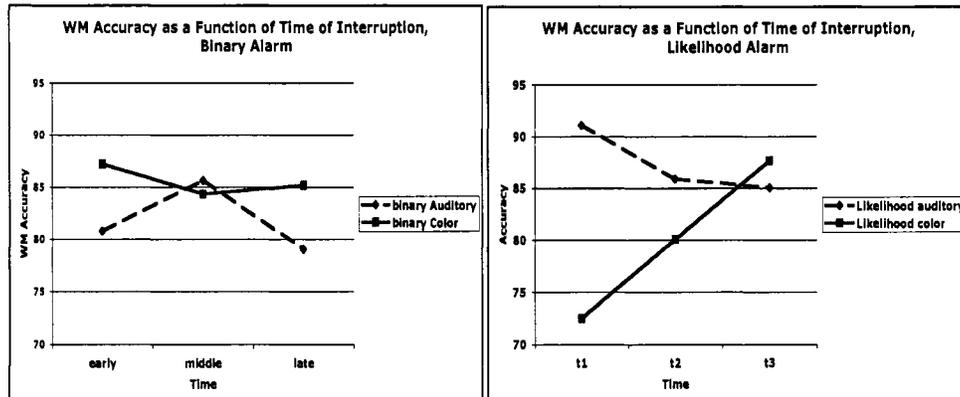
Figure 22: Working memory accuracy for each interruption time as a function of alert modality.



As shown in Figure 23, a three-way interaction between alarm type, alert modality, and time of alert presentation during the working memory task (early, middle, late) was also statistically significant, $F(1, 11) = 4.12, p. < 0.05$, suggesting that the pattern in Figure 21 is

unique to the likelihood alarm. There were no other interactions between the time interval and other factors.

Figure 23: Working memory accuracy at each interruption time for Binary Alarm (left panel) and Likelihood alarm (right panel) as a function of alert modality.



Concurrent working memory task performance was also computed for trials that were interrupted by the mid-level likelihood alert compared to trials that were interrupted by the high-level likelihood alert. There was no significant difference in accuracy between these two types of alerts (84% vs. 83% respectively). However, when mid-level likelihood alerts were compared with identical trials from the binary alert condition, participants were marginally more accurate in the binary alert condition compared to the likelihood alert condition ($M = 88\%$ vs. 84% respectively) ($F(1,11) = 3.48, p = .089$). This is contradictory with what we would expect given the longer response times during the mid-level likelihood alert. We would expect that the amber alert would signal less urgency to the participant, therefore prolonging the response times, but preserving ongoing task performance.

14. RESULTS: EXPERIMENT 2 VS. EXPERIMENT 3

Experiment 2 and Experiment 3 were identical except that Experiment 2 employed a concurrent tracking task where as Experiment 3 employed a concurrent auditory working memory task. We were interested in comparing participants' performance Experiment 2 and Experiment 3 to determine the impact of the type of concurrent task on alerted task performance in two experiments with alerts that were false-alarm prone. Note that we could not compare performance on the concurrent tasks between these experiments because they had no common metric.

14.1 Alerted Task Performance

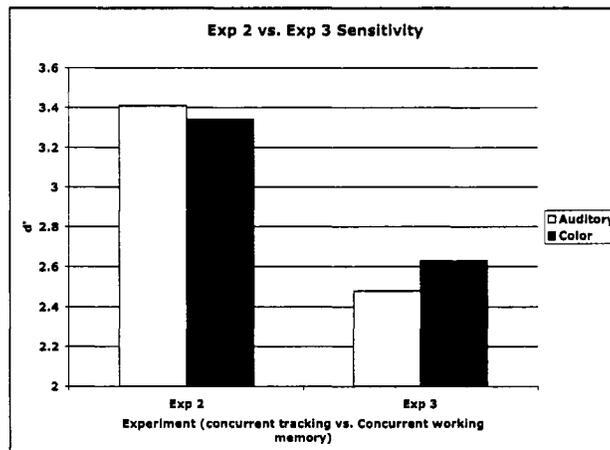
Response Time

Overall response time did not differ between Experiment 2 and Experiment 3, $F(2,24) = 1.36$, $p. > 0.10$. There were no other significant interactions between Experiment 2 and Experiment 3.

Sensitivity

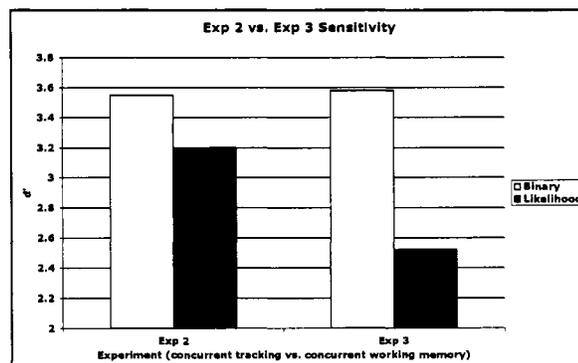
Participants were more accurate in detecting CDTI conflicts in Experiment 2, when the concurrent task was tracking, than in Experiment 3, when the concurrent task was auditory working memory ($M = 3.37$ vs. $M = 2.55$) $F(2,24) = 10.28$, $p. < 0.05$. In addition, as shown in Figure 24, modality interacted with concurrent task type (experiment) ($F(2,24) = 5.67$, $p. < 0.05$) to produce especially poor conflict detection accuracy with auditory alerts in Experiment 3 when the concurrent task was also auditory (auditory working memory).

Figure R24: Sensitivity (d') for Experiments 2 & 3 as a function of alert modality.



As shown in Figure 25, there was also a marginally significant interaction between alarm type (binary vs. likelihood) and concurrent task type (experiment) ($F(2,24) = 3.48, p > .05$), with a more severe likelihood cost with the concurrent auditory working memory task compared to the concurrent tracking task.

Figure R25: Sensitivity (d') for Experiments 2 & 3 as a function of alarm type.



15. METHODS: EXPERIMENT 4

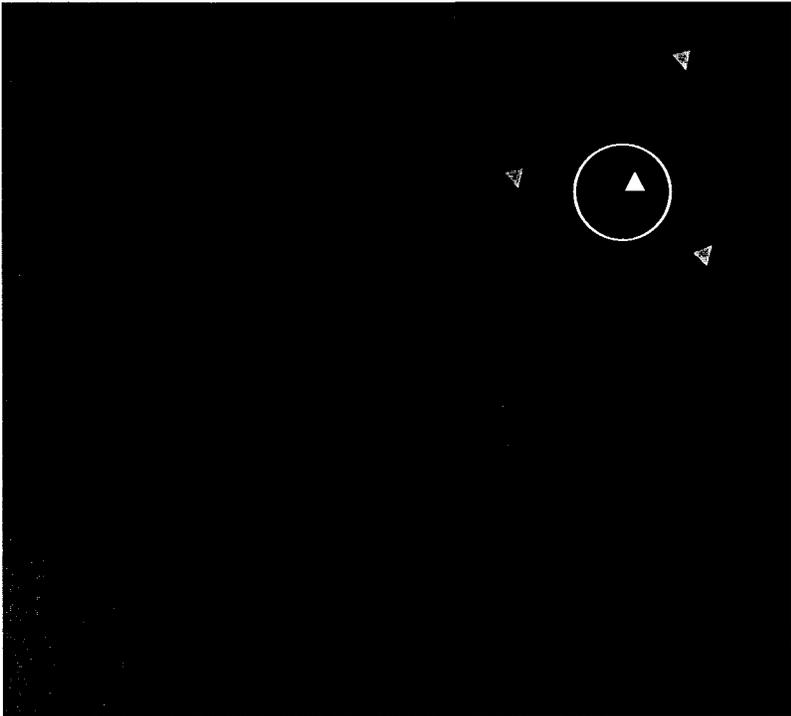
15.1 Participants

Twelve pilots from the University of Illinois Institute of Aviation were recruited to participate in the experiment. Participants ranged in age from 19 to 26 with a mean age of 22.8. Participants had normal, or corrected to normal vision. Participants had an average of 102 hours of flight experience. Participants were paid \$9 per hour for participating. Total participation time did not exceed 1.5 hours.

15.2 Procedure

Figure 26 illustrates the display used for Experiment 4. The procedure for Experiment 4 was identical to that of Experiment 3 except that the working memory number sets were presented simultaneously on screen (visually) instead of in serial order auditorally. The visual working memory stimuli were presented centrally, in the same location as the tracking task in Experiments 1 and 2 and persisted on the screen for 10 seconds and then disappeared. This 10 sec interval was equivalent to the delay between the start and the last digit presented in the auditory condition (see Figure 7). Participants then had 10 seconds to respond vocally as they did in Experiment 3.

Figure 8: Display used for Experiment 4.



16. RESULTS: EXPERIMENT 4

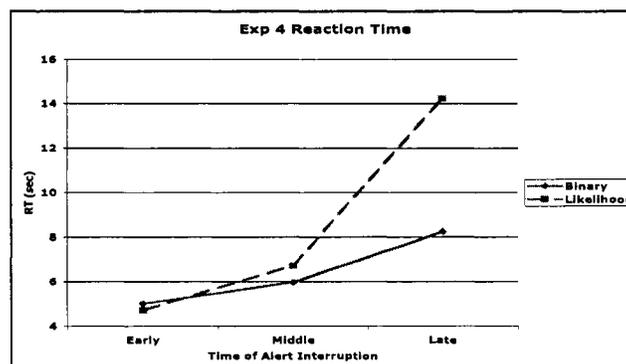
16.1 Alerted Task Performance

Response Time

There was no effect of alert modality on participants' response times to CDTI conflicts, $F(1,11) = 1.28$, $p. < 0.05$. However there was a cost in terms of response time for the likelihood alert compared to the binary alert ($M = 8.43$ vs. $M = 7.05$: $F(1,11) = 31.65$, $p. < 0.05$). Alert modality and alarm type did not interact to impact reaction time to conflicts, $F(1,11) = 3.13$, $p. > 0.10$. There were no other significant interactions.

As shown in Figure 27, when data were parsed by when in the working memory task an alert was presented (early, middle, or late), response times tended to get longer as the alert was presented later ($M = 4.86$, vs. $M = 6.35$, vs. $M = 11.24$), $F(2,24) = 36.94$, $p. < 0.01$, as had been observed in Experiment 3. In addition, presentation time of the alert interacted with alarm type ($F(2,24) = 30.62$, $p. < 0.01$) to produce an especially high cost to conflict detection time in the likelihood condition when alerts were presented near the end (late) in the working memory task, again replicating the effects in Experiment 3.

Figure 27: Response time to CDTI conflicts for each interruption time as a function of alarm type.



When response time data were compared between the mid-level likelihood alert and the high-level likelihood alert, a significant effect emerged such that participants were again slower

to respond to the mid-level alert ($M = 11.83$) compared to the high level alert ($M = 7.34$) ($F(1, 11) = 29.71, p < .05$). No significant effects emerged for events signaled by mid-level likelihood alerts and those same events signaled by a binary alert.

Sensitivity

The modality of the alert did not impact participants ability to detect CDTI conflicts ($F(1,11) = 1.19, p. > 0.10$). While alarm type (binary vs. likelihood) did not reveal a significant trend ($F(1,11) = 3.01, p > 0.10$), the non-significant trend indicated an advantage for the likelihood binary alert ($M = 3.91$) relative to the likelihood alert ($M = 3.54$). Alert modality and alarm type did not interact to impact participants accuracy in detecting CDTI conflicts, $F(1,11) = 0.50, p. > 0.10$. There were no other significant main effects or interactions.

16.2 Concurrent Task Performance

Working Memory Accuracy

Overall accuracy was quite high. There were no significant effects of alert modality ($F(1,11) = 0.78, p. > 0.10$), or alarm type ($F(1,11) = 0.90, p. > 0.10$) on concurrent working memory performance in Experiment 4. An interaction between alert modality and alarm type was also not found, $F(1,11) = 0.64, p. > 0.10$

When data were parsed based on when during the working memory task an alert was presented, there were no significant main effects of time of interruption $F(1,11) = 0.396, p > 0.05$, nor were there any significant interactions due to time of interruption. In addition, there were no differences in computation task performance when the task was interrupted by a mid-level likelihood alert, a high level likelihood alert, or the matched-trial binary alerts.

17. RESULTS: EXPERIMENT 3 VS. EXPERIMENT 4

Experiments 3 and 4 were compared to determine the impact the concurrent auditory working memory task used in Experiment 3 had on performance compared to the concurrent visual working memory task used in Experiment 4. While there were no statistical differences in response time between the experiments, conflict detection accuracy (d') was much higher in Experiment 4 compared to Experiment 3 ($M = 3.72$, vs. $M = 2.56$) $F(2,24) = 19.06$, $p. < 0.01$. In addition, task performance was better in Experiment 4 (when the task was presented visually) compared to Experiment 3 (when the task was presented auditorally and imposed working memory demands) ($M = 97.6\%$ vs. $M = 85.6\%$) $F(2,24) = 24.98$, $p. < 0.01$. There were no other significant interaction.

18. RESULTS: EXPERIMENT 2 vs. EXPERIMENT 4

Experiments 2 and 4 were compared to determine the different impact on performance generated by a concurrent visual tracking task compared to a concurrent visual working memory task (i.e. a spatial vs. verbal concurrent task, both with visual input). Responses were significantly longer in Experiment 2 compared to Experiment 4 ($M = 9.5s$ vs. $M = 7.4s$) ($F(1,11) = 15.04, p. < .05$)

19. DISCUSSION

19.1 Interpretation of Results

The current experiment was designed to evaluate six hypotheses that emerged from the collective wisdom of the prior research on attention switching, task management, modality differences and alarm/alert theory. Our conclusions regarding these hypotheses inform the applied psychology of human-system interaction in multi-task environments, and also, in part, some aspects of CDTI alert design in the cockpit. However in the latter instance, we acknowledge that some compromises were made to the true ecological nature of the CDTI, compromises we felt were necessary in order to cleanly address the scientific issues. We address these six hypotheses first, before turning to a general discussion of the relevance of the results to CDTI design.

Our first hypothesis (H1) stated that higher difficulty tracking (OT), because of its instability, would be more engaging, and thus pilots would be more reluctant to switch attention to the alert task, hence delaying response to the latter (and possibly degrading its accuracy). Across the first two experiments, there was minimal support for this hypothesis (although there was no refutation either). Responses to CDTI conflicts were not delayed by increasing tracking difficulty in either Experiment 1 or Experiment 2, and accuracy was not degraded in E1. Only in E2 was there a loss of conflict detection accuracy with higher tracking difficulty, and this effect itself was only observed with the likelihood alert. Thus pilots appear to disengage as rapidly from a difficult (unstable) concurrent task as they do from an easy unstable concurrent task, at least as this disengagement was indexed by the RT to the interrupting task (choosing the direction of turn to avoid conflict).

We also note that the fact that our participants were just as fast to respond to CDTI alerts when engaging in an unstable tracking task as a stable one could be due to participant's ability to stabilize the tracking task using peripheral vision.

Hypothesis two (H2), posits that the auditory computation task with its working memory demands would resist interruption more than the visual computation task. Such a prediction was not supported by the data. CDTI alert RT in Experiment 3 was no slower than that in Experiment

4, as would have been predicted by the hypothesis. Furthermore, had attention remained longer on the auditory working memory task (before switching), one might have predicted better performance on the less interrupted auditory computation task, than on the visual task. Instead, the effect was reversed. The auditory computational task accuracy suffered more than the visual. Therefore, we believe that a different mechanism from a strategic switch delay was responsible. Such a mechanism is probably the high interference between the working memory computations in the auditory task, and the concurrent demands of processing the traffic conflict information. Such computations were reduced, or could be more easily interleaved, with the visually presented computation task. More discussion of the role of multiple resources in accounting for this task interference will be covered below (H6).

Hypothesis 2b predicts another form of strategic influence: the rapid switch of attention from concurrent to interrupting task if the interruption occurs early in the OT sequence – to the advantage of the former, but the cost of the latter -- but the slower switch if it occurs later, as more processing has been invested in the ongoing task computations, and hence there is a greater reluctance to “leave it” until completed. Accordingly, there should be a reciprocity between CDTI response speed, and working memory computational accuracy, as the interruption time is varied. In Experiment 3, this reciprocity was nicely observed when the alert was a visual one, but not when the alert was an auditory one. Instead, with the auditory alert, both tasks appeared to suffer as the alert appeared progressively later, at a time when working memory was more heavily loaded (i.e., late in the concurrent task episode shown in Figure 6). Such an effect would seemingly reflect the heavy competition for auditory-phonetic resources between the two tasks, a competition considerably reduced when non-linguistic visual (color) alerts were employed.

Interestingly, in Experiment 4 when the computational task was visual, evidence for this between-task interference was greatly reduced. Again, later alerts led to longer switches to the alerting task reflecting a strategic delay. However this delay led to neither better nor worse performance on the computational task which, without heavy working memory demands, was performed quite well in all cases.

Our third hypothesis (H3) was that auditory alerts would pre-empt the concurrent task and consequently lead to speeded CDTI responses (and possibly higher detection accuracy to

CDTI conflicts), but at the expense of the concurrent task (increased tracking and visual computation task error). To examine this hypothesis, we considered the pattern of effects on response times to CDTI conflicts, and concurrent task error, within the three experiments that contained a visual concurrent task., and with greatest expectation for observing preemption effects when the concurrent task was tracking (as this was the concurrent task that most frequently rendered a preemption pattern in prior research; Wickens & Liu, 1988).

Partial support for the hypothesis was obtained. Across the three experiments, one experiment (Experiment 2) showed faster response times to CDTI conflicts (reduced switching time) with the auditory alerts, and no experiments showed slower responses. The two experiments (Experiments 1 and 2) that used a concurrent tracking task showed higher accuracy for CDTI conflicts during auditory alerting, and none of the three experiments showed lower accuracy for CDTI conflicts during auditory alerting. For the concurrent task however this syndrome of preemption was only partially supported. In Experiment 2, the concurrent visual tracking task did indeed degrade more when the alert was auditory than when it was visual (predicted by preemption). In Experiment 4 there was no effect, but in Experiment 1 the opposite pattern was observed (although only for the likelihood alert). That is, **both** tasks were better served by an auditory than a visual alert, an explanation more consistent with multiple resource theory (see below).

One reason the expected preemption effect (Wickens, 2003) disappeared in Experiment 1, when misses were more frequent, and false alerts were more scarce may be because the higher miss-rate degraded reliance (as predicted in hypothesis 4) and therefore forced participants to expend more visual attention in order to continuously monitor the raw data. Under such circumstances, when visual resources are more scarce, the predictions of multiple resource theory, of a benefit for auditory offload, become quite relevant (see hypothesis 6 and Wickens, Goh, Helleberg, Horrey & Talleur, 2003).

In Experiment 4, where the evidence (based on the concurrent task performance) in choosing between the two mechanisms (preemption vs. multiple resources) was ambivalent, it may well be that the two mechanisms offset each other. On the one hand, auditory preemption would degrade concurrent task performance, but on the other hand, this modality would assist the

concurrent task because of the latter's requirement for foveal vision. The discussion of multiple resource effects will be continued below within the context of hypothesis 6.

Hypothesis 4 addressed the predicted effects of alert threshold shift from Experiment 1 to Experiment 2. Here we obtained relatively solid support. As the automation False alarm rate increased, and the automation miss rate decreased from Experiment 1 to Experiment 2, two parallel predictions were offered, based on the compliance/reliance distinction of Meyer (2001). Regarding compliance, it was hypothesized that an increased alert false alarm rate would generate the so called "cry wolf" effect, leading to poorer performance on the alerting task. This hypothesis was partially supported, in that detection response times tended to increase from Experiment 1 to Experiment 2, although the interaction indicated that the increase was only present with easy tracking and the likelihood alarm. Thus pilots with a false alarm prone likelihood alert were sometimes more reluctant to disengage from their ongoing tracking task, than those with the more miss-prone alert. In contrast, conflict detection accuracy was actually improved as the threshold alert was made more sensitive from Experiment 1 to Experiment 2 (Figure R11). Since the sensitivity of the alerts themselves were virtually the same (1.33 vs. 1.53 for experiments 1 and 2 respectively), we believe this effect, reflecting a speed accuracy trade-off between experiments, can be explained by assuming that the greater delay in addressing the alerted conflict in Experiment 2 enables the potential conflict to have progressed closer to the point of closest passage and thereby makes the discrimination between safe and unsafe trajectories easier (higher sensitivity).

Regarding reliance, the compliance-reliance model predicts that a decreasing automation miss rate (concomitant with the threshold shift from E1 to E2) will avail more capacity for other tasks, and will, as a consequence, improve concurrent task performance. This effect was clearly observed in Figure R12. Any detrimental effects found for concurrent task performance due to the increased false alarm rate were clearly dominated by the benefits of increased reliance. Having relatively "miss free" automation available for CDTI alerting in Experiment 2, allowed ample attention to be devoted to tracking, to the benefit of the latter task. The importance of visual processing in underlying this effect, is signaled by the greater reliance benefit (or greater miss-prone automation cost) offered when the alert information was visual, than when it was auditory as illustrated by the interaction between modality and experiment (Figure R12).

Interestingly, this improved concurrent performance was not associated with more visual attention as indexed by scanning. Instead we assume it reflects greater general cognitive resources.

If there is high reliance in Experiment 2, compared to Experiment 1, as indexed by concurrent task performance, then this reliance effect should also be manifest in the response to “automation misses”, which should be fairly rapidly detected in Experiment 1 (when reliance is low, and the raw data are the focus of greater attention), compared to Experiment 2, when reliance is higher, automation misses are less expected, and so the raw data required to support detection of these missed conflicts receive less attention. This was the case in the binary alarm condition, with misses being detected in the binary condition of Experiment 1 over a full second faster than in Experiment 2 (compare left sides of Figure R2 and R6; difference of 1.3 seconds). One somewhat unexpected effect between Experiment 1 and Experiment 2 was the dramatic improvement in CDTI detection sensitivity that resulted, as the alert threshold was reduced, and automation misses became less likely. Since the sensitivity of the alarms themselves was essentially equivalent (1.33 v.s 1.54 for Experiments 1 & 2 respectively) , the reason for this shift remains unclear, but may pertain to the reduced visual demands of raw data monitoring in Experiment 2 in addition to the speed accuracy tradeoff mentioned earlier.

Hypothesis 5 proposed that the likelihood alarm would improve performance. Table 7 shows the pattern of “likelihood benefits” across the three dependent variables and four experiments.

Table 7: Costs and benefits of likelihood alerts: A “+” indicates support for a likelihood alert benefit, a “-” indicates support for a binary benefit, and a 0 indicates no effect. Parenthetical words signal an interaction such that the likelihood cost (benefit) is only present in the represented condition.

	Response Time	Sensitivity (d')	Concurrent Task
E1	0	0	- (aud)
E2	0	-(difficult tracking)	0
E3	-	+ (auditory)	0
E4	-	0	0

The pattern of data presented in Table 7 suggests that the greater information content of the likelihood alarm requires longer to process, and hence a greater response time (2 out of 4 experiments). However this added time will not be a concern if it buys greater accuracy or better preserves performance on the concurrent task. In Experiment 3 it is apparent that the added time does buy accuracy (at least if the alert is auditory); but in Experiment 2, the one case where there was no time penalty for the likelihood alarm, the accuracy in detecting conflicts declined (with difficult tracking).

Turning to the concurrent task, there was no evidence that the likelihood alert improved its performance (e.g., by allowing more effective or timely switching, depending upon the severity of the alert). This was somewhat surprising, in light of Woods concept of pre-attentive referencing (1995). Our further in depth analysis examined performance on those subset of conflict trials where the likelihood alert would provide its unique mid-level signal (orange or "traffic"). Across all four experiments, response time to the mid-level alert was slower than to the extreme (red, "conflict") alert. However, it is possible that this slowing resulted because those conflict trails had a separation closer to the critical boundary and therefore were harder to discriminate. Were this the case, then the corresponding trials on the binary condition should have been equally delayed. However, they were not. In all but Experiment 2, these difficult detection (mid-level) conflicts were detected more slowly with the likelihood than with the binary alert.

Finally, we can ask the detailed question of whether the likelihood alert preserved concurrent task performance better when the mid-level alert sounded. Here again, the answer was negative. No differences in tracking error were found in Experiments 1 or 2 and concurrent memory performance was actually poorer during the mid-level alert than during the corresponding binary trials.

One reason the likelihood alert did not aid task performance much in these experiments may have to do with the simplicity of the CDTI decision our pilots needed to make. In a real aviation environment, the conflict detection decision would be much more complex. It is reasonable to hypothesize that as the conflict detection decision becomes more cognitively challenging, likelihood alerts may be able to aid performance but simplifying the decision, or at

least communicating the amount of cognitive energy the pilot should invest in the alerted domain before he or she leaves the concurrent task.

Also, in contrast to the predictions of hypothesis 5b, there was not much evidence that the likelihood alarm mitigated the costs of some other effects thought to disrupt performance as the results were scrutinized in terms of interactions between alarm type and other features of increased difficulty. In Experiment 1, no interaction was found. However in Experiment 2, increasing difficulty hurt detection accuracy (d') with the likelihood alert **more**, rather than less, than with the binary alert. Furthermore, in both Experiments 3 and 4, delayed RT was found when the alert was delivered later in the task sequence as discussed above. In both experiments, this late alert time cost was actually **greater** with the likelihood than with the binary alert.

Hypothesis 6 addressed the issue of multiple resources, reflected in the general predictions that cross modal (auditory-visual) and cross code (verbal-spatial) combinations would support better performance than within modal combinations (auditory-auditory, visual-visual) and within code combinations. As noted in discussing hypothesis 3, an auditory advantage to the alerting task is predicted by both a preemption and a multiple resource (MRT) mechanism, but the observation of an auditory advantage to the visual concurrent task clearly supports MRT while refuting preemption. As noted there, data provided support for a MRT interpretation particularly with regard to conflict detection sensitivity in Experiment 1, since there was no auditory cost for the concurrent tracking task in Experiment 1. Noteworthy here is the fact that the advantage for cross-modal (AV) presentation can not simply be attributed to the greater visual scan requirements in the intra modal (VV) condition (visual alert with visual tracking), as had been the case in Wickens et al., (2003). This is because considerable attention was given in the pilot experiment, to insure that the visual color alert was equally salient to the auditory voice alert.

In addition to the results of Experiment 1, further support for the role of MRT was provided in the between-experiment analysis of Experiment 2 and 3 in which the pattern of a crossover interaction on sensitivity (Figure R18) showed that when the concurrent task (tracking) was visual (Experiment 2), the auditory alert supported more accurate detection performance than the visual, whereas when the when the concurrent task (working memory) was auditory

(Experiment 3), the pattern of interference reversed, and accuracy was better with a visual alert. Of course this between-experiment comparison confounds the processing code of the task (spatial- verbal) with modality (visual-auditory), and it may well be that the auditory cost in Experiment 3 can be attributed as much to the role of the verbal-phonetic working memory loop (processing code) in computing the digit value differences, as to the actual auditory presentation of these digits. Indeed in Experiment 4, when auditory presentation was replaced by visual presentation, the auditory cost disappeared; but was never replaced by a visual cost. Hence it is likely that both code and modality interference (Wickens, 2002) contributed to the differential pattern of task interference between Experiments 2 and 3. Final positive evidence for the role of processing code interference was the finding in the Experiment 2 – Experiment 4 comparison, when both concurrent task inputs were visual, that there was greater mutual interference with tracking (Experiment 2, both tasks spatial) than with computation in Experiment 4 (CDTI task visual, concurrent task verbal).

19.2 Practical Implications

The major practical implications of the current results are threefold. First, it appears that likelihood alarms may not always be ideal alert systems, at least in circumstances where the operator is given ample opportunity to inspect the raw data. This latter qualification is, of course, critical, since without such raw data visibility there may be major benefits to allowing the alert system to express its degree of uncertainty. In addition, the current experiment only used the three categories to express the degree of uncertainty. An additional conflict dimension that could have been employed to generate this third level of resolution is the degree of *urgency*, as perhaps signaled by the time-to-closest passage. Clearly there are many more parameters of the likelihood alarm that need to be explored, before strong conclusions regarding its operational viability can be drawn.

Second, the current results have implications regarding the modality of presenting alert information. While convention argues for heavy reliance upon the auditory modality in this regard, the current data suggest that, so long as care is taken to assure peripheral visibility, then visual alert systems may be equally effective, if not sometimes more effective (as in experiment 3), where they serve to better protect concurrent ongoing tasks, of potentially higher priority.

Third, the current data speak favorably for the designer's tendencies to shift alerting thresholds toward more false alarm prone automation. While such a shift does indeed amplify a "cry wolf" effect, in the current experiment at least, this only prolonged alert response (by about a second) without sacrificing accuracy, and the side benefit was the improved concurrent task performance, fostered by the greater reliance accompanying such a threshold shift.

REFERENCES

- Adamczyk, P.D. & B.P. Bailey. (2004). If not now, when? The effects of interruption at different moments within task execution. *Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI 2004*.
- Bailey, B.P., Konstan, J.A. & Carlis, J.V. (2001). The effects of interruptions on task performance, annoyance, and anxiety in the user interface. *Proceedings of INTERACT*, pp. 593-601.
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775-779.
- Banbury, S.P., Macken, W.J., Tremblay, S., & Jones, D.M., (2001). Auditory distraction and short-term memory: Phenomena and practical implications. *Human Factors*, 43(1), 12-29.
- Breznitz, S. (1983). *Cry-wolf: The psychology of false alarms*. Hillsdale, N.J.: Lawrence Erlbaum.
- Bustamante, E.A, Anderson, B.L., & Bliss, J.P. (2004). Effects of varying the threshold of alarm systems and task complexity on human performance and perceived workload. *Proceedings of the 48th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1948-1952).
- Chao, C., Madhavan, D., & Funk, K. (1996). Studies of cockpit tasks management errors. *International Journal of Aviation Psychology*, 6(4), 307-320.
- Cotté, N., Meyer, J., & Coughlin, J. F. (2001). Older and younger driver's reliance on collision warning systems. *Proceedings of the 45th Annual Meeting of the Human Factor Society* (pp. 277-280). Santa Monica, CA: Human Factors and Ergonomics Society.
- Crocoll, W. M. & Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. *Proceedings of the 34th Annual Meeting of the Human Factors Society* (pp. 1524-1528). Santa Monica, CA: Human Factors.

- Dixon, S.R., & Wickens, C.D. (2003). *Imperfect automation and unmanned aerial vehicle flight control* (Technical Report AHFD-03-17/MAAD-03-2). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Dixon, S.R., & Wickens, C.D. (2004). *Reliability of automated aids for unmanned aerial vehicle flight control: Evaluating a model of automated dependence in high workload* (Technical Report AHFD-04-05/MAAD-04-1). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (1999). Misuse and disuse of automated aids. *Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 339-343). Santa Monica, CA: Human Factors and Ergonomics Society.
- Funk, K. (1991) Cockpit task management: Preliminary definitions, normative theory, error taxonomy, and design recommendations. *International Journal of Aviation Psychology*, 1(4), pp(271-285).
- Gillie & Broadbent (1989). What makes disruptions disruptive? *Psychological Research*, 50, 243-250.
- Ho, C., Nikolic, M.I., & Waters, M.J. & Sarter, N.B. (2004). Not now!: Supporting interruption management by indicating the modality and urgency of pending tasks. To Appear in *Human Factors*, 46(3).
- Iani & Wickens (2004). Factors affecting task management in aviation. *Proceedings of the 48th Annual Meeting of the Human Factors Society* (pp xxx –xxx). Santa Monica, CA: Human Factors and Ergonomics Society.
- Krois, P. (2001). Alerting systems and how to address the lack of base rate information.(Unpublished manuscript.). Washington, DC: FAA.
- Kuchar, J.K. (2001). Managing uncertainty in decision –aiding and alerting system design., In *Proceedings of the 6th CNS/ATM Conference* (pp. 27-29). Taipei, Taiwan.

- Lee, J.D. & Moray, N. (1994). Trust, self-confidence, and operator's adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153-184.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80.
- Maltz, M. & Meyer, J. (2003). Use of warnings in an attentionally demanding detection task. *Human Factors*, 45(2), 217-226.
- Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. *Human Factors*, 45, 281-295.
- McFarlane, D.C., & Latorella, K.A. (2002). The score and importance of human interruption in human-computer interface design. *Human Computer Interaction*, 17(1), 1-61.
- McGowan, A., & Banbury, S. (2004). Interruption and reorientation effects of a situation awareness probe on driving hazard anticipation. *Proceedings of the 48th Annual Meeting of the Human Factor Society* (pp. 290-294). Santa Monica, CA: Human Factors and Ergonomics Society.
- Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*.
- Miller, S.L. (2002). Window of opportunity: Using the interruption lag to manage disruption in complex tasks. *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 245-249). Santa Monica, CA: Human Factors and Ergonomic Society
- Monk, C.A., (2004). The effect of frequent versus infrequent interruptions on primary task resumption. *Proceedings of the 45th Annual Meeting of the Human Factor Society* (pp. 277-280). Santa Monica, CA: Human Factors and Ergonomics Society.

- Monk, C.A., Boehm-Davis, D.A. & Trafton, G.J., (2002). The attentional costs of interrupting task performance at various stages. *Proceedings of the 46th Annual Meeting of the Human Factor Society* (pp. 1824-1828). Santa Monica, CA: Human Factors and Ergonomics Society.
- Moray, N. (2000). Are observers ever really complacent when monitoring automated systems? *Proceedings of the IEA 2000 / HFES 2000 Congress* (vol. 1, pp. 592-595). Santa Monica, CA: Human Factors and Ergonomics Society.
- Moray, N. & Rotenberg, I. (1989). Fault management in process control: eye movements and action. *Ergonomics*, 32(11), 1319-1342.
- Parasuraman, Hancock, & Olofinoboba (1997). Alarm effectiveness in driver centered collision warning system. *Ergonomics*, 39, 390-399.
- Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology*, 3(1), 1-23.
- Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, and abuse. *Human Factors*, 39(2), 230-253.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30(3), 286-297.
- Sarter, N. B. & Schroeder, B. (2001). Supporting decision-making and action selection under time pressure and uncertainty: The case of inflight icing. *Human Factors*, 43(4), 573-583.
- Sorkin, R.D., Kantowitz, B.H., & Kantowitz, S.C. (1988). Likelihood alarm displays. *Human Factors*, 30(4), 445-459.
- Spence, C., (2001). Crossmodal attentional capture: A controversy resolved? In C. L. Folk & B. S. Gibson (Eds.), *Attraction, distraction, and action: multiple perspectives on attentional capture*. Elsevier Science B.V.

- Spence, C., Nicholls, M.E.R., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics*, 63, 330-336
- St. John, M., & Manes, D.I., (2002). Making unreliable automation useful. *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 332-336). Santa Monica, CA: Human Factors and Ergonomic Society.
- Swets, J.A., (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist* 47(4), 522-532.
- Thomas, L., & Wickens, C.D., (2004). *Proceedings of the 48th Annual Meeting of the Human Factor Society* (pp. 277-280). Santa Monica, CA: Human Factors and Ergonomics Society.
- Wickens, C. D. (2000). *Imperfect and unreliable automation and its implications for attention allocation, information access and situation awareness* (Technical Report ARL-00-10/NASA-00-2). Savoy, IL: University of Illinois, Aviation Research Laboratory.
- Wickens, C. D., Gempler, K., & Morphew, M. E. (2000). Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors*, 2(2), 99-126.
- Wickens, C.D., & Hollands, J.G. (2000). Complex systems, process control, and automation. In C.D. Wickens & J.G. Hollands (Eds.), *Engineering psychology and human performance* (3rd ed.) (pp. 538-550). Prentice Hall, NJ: Upper Saddle River.
- Wickens, C.D., & Liu, Y. (1988). Codes and modalities in multiple resources: A success and a qualification. *Human Factors*, 30, 599-616.
- Woods, D.D. (1995). The alarm problem and directed attention in dynamic fault management. *Ergonomics*, 38(11), 2371-2395.
- Yeh, M. & Wickens, C. D. (2001). Display signaling in augmented reality: The effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, 43(3).

Young, S.M., & Stanton, N.A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors* 44(3), 365-375.

CURRICULUM VITAE

Angela M. Colcombe

3901 Crowwood Dr. #201
Champaign, IL, 61822
USA

Phone: (217) 766-6885

Email: acolcomb@cyrus.psych.uiuc.edu

EDUCATION

PhD (Human Factors, Psychology), May, 2006
University of Illinois, Urbana-Champaign, IL

M.A. (Visual Cognition and Human Performance, Psychology), December, 2003
University of Illinois, Urbana-Champaign, IL

B.S (Psychology/Biology), May, 1995
Northern Michigan University, Marquette, MI

RESEARCH EXPERIENCE

*Developed methodologies, collected and analyzed data, interpreted results, wrote manuscripts and/or presented results at professional conferences for following projects.

Graduate Research Assistant 2004-2006

Aviation Human Factors Laboratory,
Institute of Aviation, Department of Psychology, and The Beckman Institute
University of Illinois at Urbana-Champaign

Graduate Research Assistant 2000-2004

Visual Attention and Aging Laboratory,
Psychology Department and The Beckman Institute
University of Illinois, Urbana-Champaign

TEACHING EXPERIENCE

2001-2006 Teaching Assistant, (Supervisor, Dr. Sandra Goss-Lucas),
2006 Graduate Teaching Certificate, (Supervisor, Dr. Sandra Goss-Lucas)
2000 Teaching Assistant, (Supervisor, Dr. Thomas Srull)
2000-2006 Principles and Methods of Teaching Psychology (psych 570)

AWARDS

1999 NIMH Cognitive Psychophysiology Training Grant
2000 NIMH Cognitive Psychophysiology Training Grant
2004 Cognitive Science / Artificial Intelligence Fellowship

PUBLICATIONS

Colcombe, A.M., & Wickens, C.D. (2006) Cockpit Display of Traffic Information
Automated Conflict Alerting: Parameters to Maximize Effectiveness and
Minimize Disruption in Multiple-Task Environments. Technical Report (AHFD-05-
22/NASA-05-9), Savoy, IL. Institute of Aviation.

Kramer, A.F., Boot, W.R., McCarley, J.S., Peterson, M.S., Colcombe, A.,
Scialfa, C.T. (in press). Aging, Memory and Visual Search. *Acta-Psychologica*.

McCarley, J.S., Kramer, A.F, Colcombe, A.M., & Scialfa, C.T. (2004). Priming of pop-
out in visual search: A comparison of young and old adults. *Aging,
Neuropsychology, and Cognition*. 11(1): 80-88.

Colcombe, Angela; Kramer, Arthur F; Irwin, David E; Peterson, Mathew S; Colcombe, Stanley J; and Hahn, Sowon. (2003). Age related effects of attentional and oculomotor capture by onsets and color singletons as a function of experience. *Acta Psychologica* 113 (2): 205-225.

McCarley, J.S., Kramer, A.F., Wang, R.F., Scialfa, C.T., Colcombe, A.M., Peterson, M.S., & Irwin, D.E. (2002). How much memory does oculomotor visual search have? *Perception*, 31 (Supplement), 170.

Kramer, Arthur F; Hahn, Sowon; McAuley, Edward; Cohen, Neal J; Banich, Marie T; Harrison, Cate; Chason, Julie; Boileu, Richard A; Bardell, Lynn; Colcombe, Angela; Vakil, Eli. Exercise, aging and cognition: Healthy body, healthy mind? (Chapter). Rogers, Wendy A. (Ed); Fisk, Arthur D. (Ed); et al (2001). *Human factors interventions for the health care of older adults.* (pp. 91-120). x, 292pp.

Irwin, David E; Colcombe, Angela M; Kramer, Arthur F; Hahn, Sowon. Attentional and oculomotor capture by onset, luminance and color singletons. *Vision Research*. Vol 40(10-12) 2000, 1443-1458. Elsevier Science, England.

Kramer, Arthur F; Hahn, Sowon; Cohen, Neal J; Banich, Marie T; McAuley, Edward; Harrison, Catherine R; Chason, Julie; Vakil, Eli; Bardell, Lynn; Boileau, Richard A; Colcombe, Angela. Aging, fitness and neurocognitive function. *Nature*. Vol 400(6743). Jul 1999, 418-419. Nature Publishing Group, United Kingdom.