

PILOTS' CONFLICT DETECTION WITH IMPERFECT CONFLICT ALERTING SYSTEM
FOR THE COCKPIT DISPLAY OF TRAFFIC INFORMATION

Xidong Xu, Christopher D. Wickens, and Esa M. Rantanen
University of Illinois at Urbana-Champaign
Savoy, Illinois

Twenty-four pilots viewed dynamic encounters between the pilot's "ownship" and an intruder aircraft on a 2-D simulated Cockpit Display of Traffic Information (CDTI) and estimated the point and time of closest approach. A three-level alert system provided a correct categorical estimate of the projected miss distance (MD) on 83% of the trials. The remaining 17% of alerts incorrectly predicted MD. The data of these pilots were compared with a matched "baseline" pilots, who viewed identical trials without the aid of automated alerts. Roughly half the pilots depended on and benefited from this automation, and others did not. Those who benefited did so when problems were difficult but not when they were easy. Furthermore, automation benefits were observed only when automation was correct, but automation costs were not observed when it was in error. While assisting miss distance prediction, the automation led to an underestimate of the time remaining till the point of closest approach.

Introduction

Cockpit Display of Traffic Information (CDTI) will play a key role in a new flight environment known as free flight, allowing pilots to detect and avoid potential conflict by providing graphic information regarding nearby traffic's locations, speeds, altitudes, and other information relative to the ownship (Johnson, Battiste, & Bochow, 1999; Wickens, Helleberg, & Xu, 2002). Airborne conflict detection is a cognitively demanding task, and consequently automated aids have been invoked to assist pilots in their new charge. However, future flight paths are inherently uncertain due to a number of factors such as wind shift, pilots' intentions to change flight plans, and look-ahead time, making perfect predictions impossible (e.g., Kuchar, 2001). These uncertainty factors may lead to two types of errors in automated conflict alerting: misses (no alert of real conflict) and false alarms (safe separations treated as conflicts). In addition to the safety consequences that may result from pilots' over-trusting or over-depending on these erroneous automation outcomes, both high false alarm rates and high miss rates may cause operators to mistrust the system, which may in turn cause under-use (or under-dependence) and even disuse of the system (Parasuraman & Riley, 1997). The implications of this unreliability of automation in conflict detection are of particular interest to us.

In general, correctly functioning automation tends to improve overall system performance relative to unaided performance (Dixon, Wickens, & Chang, in press; Metzger & Parasuraman, 2005; Yeh & Wickens, 2001). However, automation benefits may not always be realized if the manually performed task had been easy (Rovira & Parasuraman, 2002), whereas automation benefits can be substantial when

tasks are difficult in their manual form (Dixon & Wickens, 2004; Maltz & Shinar, 2003, Wickens & Dixon, 2005). On the other hand, costs of inaccurate automation may be larger for difficult tasks than for easy tasks (Dixon & Wickens, 2004; Maltz & Shinar, 2003; Wickens, Gempler, & Morphew, 2000). The costs and benefits can easily be interpreted with the mediating concept of automation *dependence*. As tasks become more difficult, users become more dependent on automation to assist them, which will provide greater benefits when the automation is correct, but greater costs when it "fails" due to reduced situation awareness and/or skill degradation resulting from complacency (Parasuraman, Sheridan, & Wickens, 2000). One noticeable phenomenon is that when reliability is above 70%–75%, there are benefits but no costs relative to manual performance, especially when the task is difficult (e.g., Wickens & Dixon, 2005; Maltz & Meyer, 2003).

It appears that only two experiments have examined the issue of human responses to imperfect automation in aviation conflict detection and avoidance, by Metzger and Parasuraman (2005) and Wickens et al. (2000). The findings of Metzger and Parasuraman (conflict detection in air traffic control) and Wickens et al. (conflict avoidance using a CDTI) collectively show that correct automation is beneficial to performance and inaccurate automation poses costs, consistent with the general pattern found in other studies. The results are also in agreement with the general finding that costs and benefits are more likely to emerge for difficult (vs. easy) task, which would more likely make people depend on automation.

The goal of the present study focused on how correct and erroneous predictions of an imperfect automation alert affected performance in relation to the unaided

baseline performance reported in Xu, Rantanen, and Wickens (2004), and how the effect of automation reliability was modulated by task difficulty. Based on the literature reviewed above, we formulated several hypotheses: (1) that conflict detection performance using a CDTI with an imperfect automated alerting system (83% reliable in the current study) predicting the miss distance (MD) at the closest point of approach (CPA) between the ownship and an intruder would be better than unaided performance (Wickens & Dixon, 2005); (2a) that correct automation with a valid MD alert would improve performance and (2b) error automation with invalid MD alert would hinder performance relative to manual performance on equivalent difficulty (Metzger & Parasuraman, 2005); (3a) that increasing trial difficulty would amplify the effect of reliability as mediated by increased dependence; that is, for correct automation, automation would provide greater performance *improvement* relative to manual performance for hard trials than for easy trials, and (3b) for automation errors, automation would induce greater performance *costs* relative to manual performance for hard trials than for easy trials (Dixon & Wickens, 2004).

Method

Participants

Twenty-four pilots (22 male and two female; age ranging between 18-25 years, with a mean of 19.8 years) different from the baseline study reported elsewhere (Xu, Rantanen, et al., 2004) were recruited from the Institute of Aviation, University of Illinois at Urbana-Champaign.

Simulation and Task

The CDTI depicted ownship and intruder in a map (top-down) view (see Figure 1). The display represented ownship by a white triangle and the intruder by a solid circle in cyan, yellow, or red, depending on the MD alert level. Ownship icon was positioned in the center of the display throughout the whole experiment, thus yielding an egocentric view of the traffic situation, where the ownship icon appeared to be stationary to the participant. The ownship and the intruder were flying at the same altitude on straight converging courses and at constant but not necessarily same speeds. At the start of a trial, a conflict predictor provided a three-level MD alert (no alert if $MD > 3.5$ nm; low level alert if $1.5 \text{ nm} < MD < 3.5$ nm; and high level alert if $MD < 1.5$ nm). The three levels of MD alert were indicated by different colors of the intruder icon, along with different verbal warnings (cyan and no verbal warning for no alert, yellow and “traffic traffic” for low level alert, and red and “con-

flict conflict” for high level alert). The intruder icon retained the color throughout a trial and the verbal warning was given once at the beginning of a trial. To simulate a less than perfectly reliable predictor, on one in every six trials, the automation provided erroneous prediction of MD, indicating MD that was in a greater (a miss) or smaller separation (a false alarm) category than the true value.

Participants individually observed the development of a conflict scenario for 15 sec, after which the scenario froze. They were then required to mentally extrapolate the development of the scenario, press a key when they estimated that the CPA was reached, thereby providing the estimate accuracy of time to CPA (TCPA), and move the cursor to a location that they believed was the CPA, thus providing the estimate accuracy of MD. Pilots were instructed that when the MD alert was correct, they were supposed to take advantage of it. However, when they believed that the predictor provided invalid MD prediction, the pilots were asked to ignore it and make their estimations based on their own judgments.

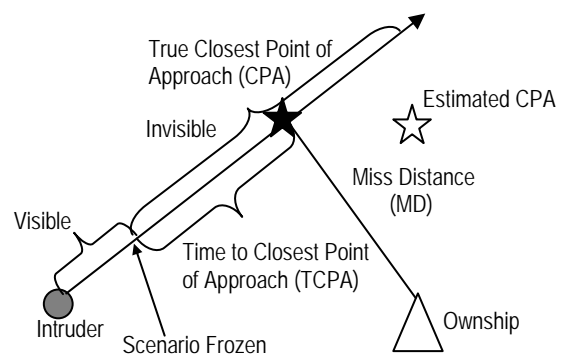


Figure 1. Schematic illustration of key components of the experimental paradigm.

Experimental Design and Procedure

This experiment employed a repeated measures design. However, the data for pilots in this experiment using automation were statistically compared with matched data from the “baseline” study, on identical conflict trials, performed without the aid of automation, as reported in Xu, Rantanen, et al. (2004). The trials with differing geometries in the baseline experiment that produced the easy and hard trials were randomly chosen to create an independent variable of task difficulty for the present experiment, which was varied within subjects. Trial difficulty was inferred from pilot performance, with smaller errors indicating easy trials and greater errors hard trials (see Xu, Wickens, & Rantanen, 2004 for full details).

The second independent variable was automation correctness (error vs. correct). Each participant received 60 correct automation trials and 12 automation error trials (reliability of 83%), the latter equally representing misses and false alarms, of large and small magnitude. The 24 pilots were chosen such that their flying experience was roughly equal to that of the corresponding pilots in the baseline experiment. The 72 trials were quasi-randomly presented to the pilot. The automation error trials were in turn quasi-randomly distributed within the total 72 trials.

The participants were explicitly told that the MD predictor would not be 100% reliable. They performed ten practice trials, with a valid predictor for the first six trials and an invalid predictor for the remaining four trials, being informed explicitly of the invalidity of the last four trials. Then they participated in one experimental session to complete two blocks of 36 trials each for one to two hours in total.

Dependent Measures

The dependant variables were absolute and signed MD and TCPA estimate errors, derived by subtracting the true values from the estimated values. Absolute errors would reveal the estimation accuracy, and signed errors would reveal the estimation directions (under- or overestimate), an indication of biases. Our attention was focused on the MD measures as pilot-estimated MD represented the most safety-critical aspect of the pilot's assessment of conflict risk.

Results

Data Reduction and Analysis

A good measure of automation dependence is the difference in performance between conditions of error and correct automation (Maltz & Shinar, 2003), with a large difference being indicative of heavy dependence. We measured automation dependence by the difference in absolute MD estimate error between the automation error trials and the correct automation trials, given that only MD prediction was automated, as well as the fact that MD is the primary measure of conflict risk. The difference was calculated separately for each individual pilot, thus yielding two levels of automation dependence (light and heavy) for the 24 pilots using a median-split method. The light dependence pilots mostly encountered easy trials, whereas the heavy dependence pilots mostly had hard trials.

Analyses for Heavy Dependence Group

Hypothesis 1 was tested by two-sample t-tests for

means and hypotheses 2 and 3 by 2×2 ANOVAs.

Overall Effect of Automation (Hypothesis 1). Absolute MD estimate error was .13 nm smaller in the current experiment ($M = .33$ nm) than in the corresponding trials collected in the baseline experiment ($M = .46$ nm), $t(22) = -1.83$, $p = .04$, suggesting that the automated alerts used here, even though imperfect, nonetheless benefited MD estimation. However, absolute TCPA estimate error did not differ significantly between the automation and manual (baseline) groups, $t(22) = .63$, $p = .27$.

MD Estimate Error (Hypotheses 2 & 3). Figure 2 presents the MD error for the baseline and automation experiments. A 2 (automation vs. manual baseline) $\times 2$ (easy vs. hard conflict problems) mixed ANOVA for the automation error trials and the corresponding baseline (manual) trials (the left side of Figure 2) revealed that absolute MD estimate error did not significantly differ between the two experiments, $F(1, 22) = .56$, $p = .46$, and performance on hard trials was poorer than on easy trials in both experiments, $F(1, 22) = 17.40$, $p < .0001$. Furthermore, the same ANOVA revealed that the difference in performance between easy and hard trials in the present experiment was not reduced compared to that in the baseline experiment, since the interaction between experimental condition and task difficulty was not significant, $F(1, 22) = .31$, $p = .59$.

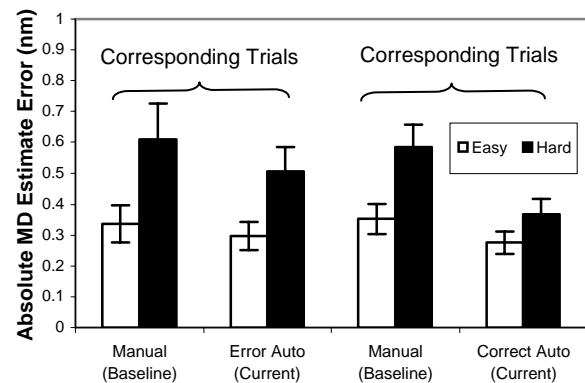


Figure 2. Absolute MD estimate errors for heavy dependence group by automation correctness and task difficulty, and the corresponding baseline trials.

In contrast, a 2 (automation vs. manual) $\times 2$ (easy vs. hard) mixed ANOVA for the correct automation trials and the corresponding baseline trials (the right side of Figure 2) revealed that performance was better (smaller error) than in the baseline experiment, $F(1, 22) = 4.19$, $p = .053$, and performance on easy trials was better than on hard trials, $F(1, 22) = 36.73$,

$p < .0001$. Most importantly, the difference in performance between the easy and hard trials in the current experiment was reduced compared to that in the baseline experiment, indicated by a significant interaction between experimental condition and task difficulty, $F(1, 22) = 6.89, p = .015$.

Analysis on *signed* MD estimate error suggests that there was a tendency for the MD to be less underestimated in the current experiment compared to the baseline experiment, especially when the automation was correct and the task was hard (see Xu, Wickens, et al., 2004 for detailed analysis). Therefore, the automation moved the *signed* estimates closer to the true value, reducing a conservative bias to underestimate MD that had been observed in the baseline study (see Xu, Rantanen, et al., 2004)

TCPA Estimate Error (Hypotheses 2 & 3). The results of a 2 (automation vs. manual) \times 2 (easy vs. hard) mixed ANOVA revealed that when the automation was present and in *error* (left half of Figure 3), absolute TCPA (time) estimate error did not differ significantly from that in the baseline experiment, $F(1, 22) = 2.35, p = .14$; and performance on the hard trials was constantly poorer than on the easy ones in both experiments, $F(1, 22) = 28.86, p < .0001$. There was no significant interaction between the two factors $F(1, 22) = .40, p = .53$.

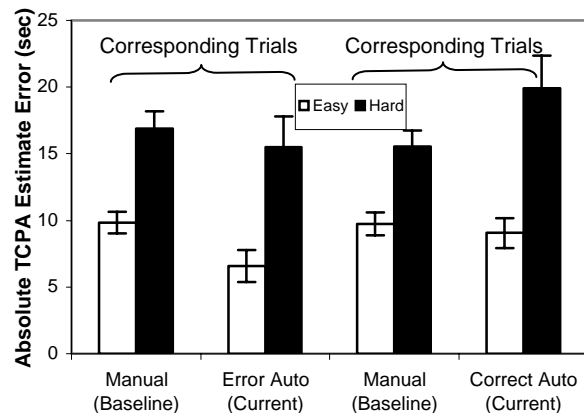


Figure 3. Absolute TCPA estimate errors for heavy dependence group by automation correctness and task difficulty, and the corresponding baseline trials.

However, when the automation was *correct* (right half of Figure 3), the increase in task difficulty imposed greater cost to performance than in the baseline experiment as indicated by greater absolute TCPA estimate error. This trend was confirmed by the results of a 2 (automation vs. manual) \times 2 (easy vs. hard) mixed ANOVA, which revealed that the hard trials were still harder than the easy ones in the cur-

rent experiment, $F(1, 22) = 69.32, p < .0001$, but the significant interaction between experimental condition and task difficulty suggests that the hard trials (with automation) in the current experiment induced greater TCPA (time) estimation error than the corresponding hard trials (without automation) in the baseline experiment, $F(1, 22) = 6.37, p = .019$.

Similar to the analyses on signed MD estimate error, we also looked at the data on *signed* TCPA estimate error (estimating conflict to be too early or too late) in both experiments. The analyses revealed that for both the correct and error automation trials, TCPA was more underestimated than their baseline counterparts, (closest passage estimated to arrive sooner than it actually would); this bias was amplified on hard trials, in both experiments (see Xu, Wickens, et al., 2004 for more detailed analysis).

Analyses for Light Dependence Group

Analyses were performed for the light dependence group in the same way as those for the heavy dependence group and the results show that none of the hypotheses were supported for the light dependence group, $F_s < 1$ and $t_s < 1$. This pattern of results thus suggests that the light dependence pilots did not use the automation, and hence were unaffected by its properties. Moreover, absolute MD estimate error was greater on the hard trials than on the easy trials for both the correct automation trials and the corresponding baseline trials, $F(1, 22) = 16.36, p = .001$, and the difference between hard and easy trials was not significantly reduced by automation, $F(1, 22) = .047, p = .83$, suggesting that these pilots should have used (but did not) automation for support.

Discussion

First, we had not originally anticipated the wide range of automation dependence between participants. Given such a range, it made sense to focus our hypothesis testing regarding automation properties primarily upon those who depended on automation in the first place, since those who did not would be expected to show generally null results of automation correctness (and indeed they did). Because those low dependence pilots were people who were more likely paired with those pilots who encountered easier problems in the baseline experiment, they also generally received easier problems in the current experiment. It appears that those low dependence pilots did not feel the need to obtain assistance from the automation, presumably because the task was relatively easy, although our analysis revealed that they should still have used it to improve performance. In contrast,

since the high dependence pilots received the more difficult trials, it might have appeared to be an advisable strategy to depend on the automation to enhance performance. Indeed they generally were found to benefit from automation regarding the most critical safety-relevant or risk measure of conflict understanding, the estimation of MD at the closest point of approach. Performance of these pilots was better than that of their demographically matched counterparts in the baseline experiment, facing problems of equivalent difficulty but unaided, thus, supporting Hypothesis 1. Importantly, the data show that with an error rate of 17% (83% reliability), pilots clearly benefited from imperfect automation, a data point that adds to the general conclusion that imperfect automation above a 70-75% rate is better than no automation at all when workload is high and the task is difficult (Wickens & Dixon, 2005).

The analysis examining Hypothesis 2 revealed, as expected, that benefits were only realized when automation was correct and not when it was in error (thus supporting Hypothesis 2a; Figure 2). However, the results were a little surprising in that even on the automation error trials performance was no worse than its level had been in the baseline experiment, and sometimes showed a hint of being better (thus refuting Hypothesis 2b). That is, unlike other findings, erroneous automation did not yield a “complacency cost” of over-dependence, corresponding to an automation-induced beta shift (e.g., Maltz & Shinar, 2003; Metzger & Parasuraman, 2005; Yeh & Wickens, 2001). One partial explanation is that pilots were clearly pre-warned of the less-than perfect characteristics, and so were presumably not “caught” by a first failure effect, which is typically used to document the effect of over-trust, over-dependence, or “complacency” (e.g., Yeh & Wickens, 2001).

How did the high dependence pilots show a benefit from imperfect automation when it was correct, but no cost when it was wrong? Part of the answer may be because the pilots’ response (positioning the cursor on the location of the projected CPA) was different from the actual guidance given by the automation predicted MD. In interpreting our results, we assume that when the high- and low-level alert appeared, pilots invested a high level of perceptual and cognitive processing of the raw data—a careful inspection—in order to most accurately estimate the CPA. This effort investment was greater than that for corresponding pilots in the baseline experiment, who did not receive the alert. Such behavior would lead to enhanced accuracy even when the alert was incorrect. When the alert was “silent” in contrast, pilots might have maintained an equivalent level of inspection to

their manual baseline counterparts.

Another, parallel way of accounting for the data is to assume an overall improvement in performance of the current experiment versus the baseline experiment, perhaps due to a motivational increase from having the automation available (e.g., Beck, Dzinodlet, Pierce, & Piatt, 2003). Within the overall improved performance, the cost-benefit differences associated with automation error versus correct still existed (at least on the difficult problems; see Figure 2). However, any cost for error automation was then entirely offset by the overall benefit of improved motivation and performance, particularly when the alert sounded, as described above, triggering a closer inspection of the raw data.

The finding that automation benefits emerged on high difficulty trials (thus supporting Hypothesis 3a; Figure 3) is a familiar and expected one (e.g., Dixon & Wickens, 2003, 2004; Maltz & Shinar, 2003). It is also important to note that a major feature of the high difficulty was the long distance to the closest point of approach, creating a lengthening of space over which projection must take place, that would be typical as we extrapolate the current results to the more strategic uses of the CDTI that are envisioned (e.g., 2-4 minute look-ahead time). In such a case, pilots would either have to project across a larger region of the display or if the display scale were minified, they would have to project across a slower velocity symbol movement, a prediction that is also more difficulty (Xu, Rantanen, et al., 2004) and so, again, would be likely to benefit from imperfect automation.

Another finding that was not anticipated was the distance-time estimation accuracy trade-off that was produced by automation. That is, while automation appeared to improve the accuracy of performance on the most critical task associated with conflict estimation—the estimation of miss distance at the CPA— it actually disrupted the accuracy of estimating the time till that CPA would occur. Why this occurred may be accounted for by a resource trade-off—the requirement to process both the automated alert and the raw visual data for miss distance required more resources. Such resources were diverted from the time estimation process, which was itself resource limited (Zakay, Block, & Tsal, 1999). Given then that time would be more poorly estimated as a consequence of resource diversion, pilots adopted a “conservative strategy” to underestimate that time; that is, to give themselves less time available than they really have. In conclusion, the results have clearly illustrated the benefits that can be provided by even imperfect or “unreliable” CDTI alerting, at least given the rela-

tively high reliability level about 80%. Such benefits –without costs – are, we believe, the result of three factors: (1) Raw data were available to be inspected; (2) pilots were calibrated to the approximate reliability level, and (3) a three-level alert was employed. We might project that increases in multi-task workload to a level more typical of the cockpit might amplify the benefits, just as decreasing the automation error rate would have had the same effects. However, it is possible that these two changes, while amplifying the benefits of correct automation, may have led to the emergence of costs on automation-error trials. Finally, caution needs to be exercised when generalizing the results here regarding the effects of automation unreliability to the real world situations, where the conflict base rate is much lower than in the present study.

Acknowledgements and Authors' Note

This research was supported in part by a grant from NASA Ames Research Center (NASA NAG 2-1535). Dr. David C. Foyle was the technical monitor. Views expressed here are those of the authors and do not necessarily reflect the views of NASA. The authors thank Ron Carbonari for the programming support. The first author is now at San Jose State University Foundation at NASA Ames Research Center, CA.

References

- Beck, H. P., Dzindolet, M. T., Pierce, L. G., & Piatt, N. (2003). Looking forward: A simulation of decision aids in tomorrow's classroom. *Proc. HFES 47th Annual Meeting* (pp. 330-334). Santa Monica, CA: HFES.
- Dixon, S. R., Wickens, C. D., & Chang, D. (in press). Mission control of multiple unmanned aerial vehicles: a workload analysis. *Human Factors*.
- Dixon, S. R., & Wickens, C. D. (2004). *Reliability in Automated Aids for Unmanned Aerial Vehicle Flight Control: Evaluating a Model of Automation Dependence in High Workload* (AHFD-04-05/MAAD-04-1). Savoy, IL: UIUC, AHFD.
- Johnson, W. W., Battiste, V., & Bochow, S. H. (1999). A cockpit display designed to enable limited flight deck separation responsibility. *Proc. 1999 World Aviation Conference*. Warrendale, PA: Society of Automotive Engineers.
- Kuchar, J. K. (2001, March). *Managing uncertainty in decision-aiding and alerting system design*. Paper presented at the 6th CNS/ATM Conference, Taipei, Taiwan.
- Maltz, M., & Meyer, J. (2003). Use of warnings in an attentionally demanding detection task. *Human Factors*, 43(2), 217-226.
- Maltz, M., & Shinar, D. (2003). New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors*, 45(2), 281-295.
- Metzger, U., & Parasuraman, R. (in press). Automation in future air traffic management: Effects of reliable and imperfect conflict detection aids on controller performance and mental workload. *Human Factors*.
- 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.
- Rovira, E., & Parasuraman, R. (2002). *Sensor to shooter: Task development and empirical evaluation of the effects of automation unreliability*. Paper presented at the Annual Midyear Symposium of the APA, Division 10 (Military Psychology) and 21 (Engineering Psychology). Ft. Belvoir, VA.
- 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., Helleberg, J., & Xu, X. (2002). Pilot maneuver choice and workload in free flight. *Human Factors*, 44(2), 171-188.
- Wickens, C. D., & Dixon, S. R. (2005). Is there a magic number 7: the benefits of imperfect automation. AHFD-05-01/MAAD 05-01. Savoy, IL: UIUC, AHFD.
- Xu, X., Rantanen, E. M., & Wickens, C. D. (2004). *Estimation of conflict risk using cockpit displays of traffic information* (AHFD-04-11/FAA-04-4). Savoy, IL: UIUC, AHFD.
- Xu, X., Wickens, C. D., & Rantanen, E. M. (2004). *Imperfect conflict alerting systems for the cockpit display of traffic information* (AHFD-04-8/NASA-04-2). Savoy, IL: UIUC, AHFD.
- 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), 355-365.
- Zakay, D., Block, R. A., & Tsal, Y. (1999). Prospective duration estimation and performance. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII—Cognitive regulation of performance: Interaction of theory and application* (pp. 557-580). Cambridge, MA: The MIT Press.