Separation Management: Automation Reliability Meta-Analysis and Conflict Probe Reliability Analysis

Jonathan R. Rein, PhD, Spectrum Software Technology, Inc.
Anthony J. Masalonis, PhD, Spectrum Software Technology, Inc.
James Messina, PhD, Behavioral Science Associates, Inc.
Ben Willems, FAA Human Factors Branch

November 2012

DOT/FAA/TC-TN-12/65

This document is available to the public through the National Technical Information Service (NTIS), Alexandria, VA 22312. A copy is retained for reference at the William J. Hughes Technical Center Library.

U.S. Department of Transportation
Federal Aviation Administration

William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. This document does not constitute Federal Aviation Administration (FAA) certification policy. Consult your local FAA aircraft certification office as to its use.

This report is available at the FAA William J. Hughes Technical Center's full-text Technical Reports Web site: http://actlibrary.tc.faa.gov in Adobe® Acrobat® portable document format (PDF).

1. Report No. DOT/FAA/TC-TN-12/65	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Separation Management: Automation Reliability Meta-Analysis and Conflict Probe Reliability Analysis		5. Report Date November 2012	
		6. Performing Organization Code ANG-E25	
7. Author(s) Jonathan R. Rein, Ph.D., Spectrum Software Technology, Inc. Anthony J. Masalonis, Ph.D., Spectrum Software Technology, Inc. James Messina, Ph.D., Behavioral Science Associates, Inc. Ben Willems, FAA Human Factors Branch		8. Performing Organization Report No. DOT/FAA/TC-TN-12/65	
9. Performing Organization Name and Address Federal Aviation Administration Human Factors Branch William J. Hughes Technical Center Atlantic City International Airport, NJ 08405		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No.	
12. Sponsoring Agency Name and Address Federal Aviation Administration 800 Independence Avenue, S.W. Washington, DC 20591		13. Type of Report and Period Covered Technical Note	
		14. Sponsoring Agency Code AJM-121	

15. Supplementary Notes

16. Abstract

Objective: The purpose of this study is to establish whether a valid performance criterion exists to determine the acceptability of the En Route Automation Modernization (ERAM) Conflict Probe's conflict-detection accuracy and to evaluate observed accuracy against this criterion. **Background:** The Conflict Probe can exhibit very high or very low accuracy depending on the analysis technique. It is necessary to establish both an empirically backed criterion for accuracy and the appropriate accuracy analyses and metrics. **Method:** A meta-analysis was conducted on Human Factors automation reliability literature; an additional analysis was done on the results of probe reliability studies by the Federal Aviation Administration (FAA) Concept Analysis branch to derive various accuracy metrics. The results were compared to determine the acceptability of the conflict probe accuracy. **Results:** The meta-analysis produced an estimated criterion of 65% correct responses for automation to improve performance, but this estimate is subject to a broad confidence interval due to variability in the source data from the literature. The probe performance exceeded the 65% value when giving credit for all correct rejections, but it fell short when not giving credit for correct rejections. Another metric, Positive Predictive Value (PPV, the percent of alerts that are valid), is operationally meaningful and its values demonstrated large accuracy improvements over baseline with the FAA Concept Analysis' parametric adjustments, but a PPV cutoff criterion could not be established from the meta-analysis. **Conclusion:** The present results provide insight on several fronts, but operational input is essential to determine (1) a more justifiable air traffic control-specific accuracy criterion and (2) which aircraft encounters are appropriate to include in a test set for accuracy assessment. Automation responses should only increase the value of an accuracy metric to the extent that the responses add informational value for the controller. **Applications:** The present results will be used in the design of an evaluation to derive a set of operationally meaningful aircraft encounters. Different measures of accuracy, such as PPV, also merit further exploration.

17. Key Words Automation Reliability En Route Automation Modernization Human-Automation Performance Meta-Analysis Positive Predictive Value Signal Detection Theory		18. Distribution Statement This document is available to the public through the National Technical Information Service, Alexandria, Virginia, 22312. A copy is retained for reference at the William J. Hughes Technical Center Library.	
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 30	22. Price
Form DOT F 1700.7 (8-72)	**Reproduction of completed page authorized**		

THIS PAGE IS BLANK INTENTIONALLY.

Table of Contents

List of Illustrations

Acknowledgments

THIS PAGE IS BLANK INTENTIONALLY.

Executive Summary

The purpose of this study was to determine empirically a level of conflict detection accuracy for the En Route Automation Modernization (ERAM) Conflict Probe that would be acceptable for display on the radar side (R-side) and to assess the probe's observed accuracy in relation to that criterion. We performed a series of analyses to assess two questions: (1) Is it possible, based on the existing human factors literature, to define a minimum level of decision-support automation accuracy required to improve joint human-automation performance? (2) How well (according to appropriate accuracy metrics) can the Conflict Probe function of ERAM predict aircraft-to-aircraft conflicts?

For the first question, we conducted a meta-analysis of 12 studies from the human factors literature, studying the effect of imperfect automation on system performance. This analysis quantified the relationship between several accuracy metrics and the corresponding automation effect size (i.e., performance of a human-automation system, as compared to the human operating without an automation aid). The automation reliability metric (overall percent correct) exhibited a statistically significant positive relationship with effect size. We found a "crossover point" of 0.65 reliability, with automation accuracy greater than this value likely resulting in performance benefits. This is similar to the value obtained by Wickens and Dixon (2007), using a different methodology. Noise in the data led to a large amount of uncertainty in our estimate, however, with the 95% confidence interval ranging from 0.39 to 0.72. The other metrics explored were the basic hit and false alarm rates from signal detection theory (SDT), and Positive Predictive Value (PPV), an operationally meaningful metric commonly used in clinical work which indicates what proportion of reported alerts are valid. We did not find statistically reliable relationships between these metrics and the performance benefit realized from the automation.

For the second question, we computed the Conflict Probe's prediction accuracy on the same metrics via additional analyses based on the results of engineering analyses by Crowell, Fabian, Young, Musialek, and Paglione (2011) and Crowell and Young (2012). The values found in this analysis were compared with the results of our meta-analysis. The results indicated that when Conflict Probe performance on the reliability metric was estimated liberally—that is, when a larger set of correct rejections (CRs) was included in the input data—the automation performance exceeded the upper bound of the confidence interval for the crossover point, suggesting that it may be sufficiently accurate to improve Air Traffic Control (ATC) performance. However, when the strictest criterion was adopted and no CRs were included in the analysis, the reliability fell short of the lower bound of the confidence interval for the crossover point. For other accuracy metrics, such as hit and false alarm rates and PPV, the data from the studies reviewed for the meta-analysis did not provide sufficient statistical reliability to determine the acceptability of the Conflict Probe performance.

In summary, the present analyses lay the groundwork toward answering the stated research questions, but could not definitively establish the performance of proposed Conflict Probe algorithms, nor could they identify a clear accuracy criterion for operational acceptance. Further study is recommended (a) to determine, via operational ATC input, an appropriate cutoff for including aircraft pairs in engineering analyses so that the computed CR rate, and resulting reliability and related metrics, are neither too high nor too low to characterize performance in an operationally meaningful way; (b) to examine Conflict Probe false alarms in detail to determine whether the false alarms that are decreasing the values on the metrics are, in fact, operationally acceptable; and (c) to identify the effect of automation reliability, hit rate, false alarm rate, and PPV on performance in the specific context of Conflict Probe. Operational input is key to answering all these questions more definitively. For example, automation responses should only increase the value of an accuracy metric to the extent

that the responses add informational value for the controller and should only decrease the accuracy metric to the extent that they risk impairing automation-assisted controller performance. The present results will, therefore, be used as input to the design of evaluations to study these important questions in depth.

1. INTRODUCTION

The Human Factors Branch (ANG-E25) at the Federal Aviation Administration (FAA) William J. Hughes Technical Center (WJHTC) conducts human-in-the-loop (HITL) simulations to assess alternative en route system design elements as part of the Separation Management project. The results of these HITL simulations will be used to develop requirements for the En Route Automation Modernization (ERAM) system. Two HITL simulations have been conducted so far. The first HITL was conducted to investigate the effect of variable separation requirements within a sector and among aircraft, as well as a variety of automation tools to aid controllers' implementation of variable requirements. The second HITL was conducted to evaluate issues that include (a) the location and format of the Conflict Probe notification on the controller workstation, (b) the replacement of the current radar (R-side) and/or data (D-side) controller workstation display with a 30-inch commercial-off-the-shelf (COTS) monitor, and (c) alternative pointing devices to the existing controller workstation trackball.

The Human Factors branch is now planning to conduct a third HITL in the Separation Management series. We conducted two related data analysis tasks, and integrated the results of the two sets of analyses, to verify the need for the third HITL simulation and for preliminary research to better characterize the current accuracy of the Conflict Probe. One analysis was a more rigorous version of the informal meta-analysis conducted by Wickens and Dixon (2007) to determine a cutoff for the *reliability* that automated decision-support tools must exceed to improve the performance of the joint human-automation system. The other analysis was conducted to assess the current state of the maximum achievable reliability of the ERAM Conflict Probe. The results of the analyses were then compared to determine the acceptability of the current Conflict Probe accuracy.

In the current (HOST) implementation of the en route controller workstation, two automation features signal potential conflicts to controllers: Conflict Alert and Conflict Probe. Conflict Alert uses the momentary speed and heading of aircraft to predict losses of separation within approximately two to three minutes. When a loss of separation is predicted, the data blocks of the relevant aircraft flash and information about the predicted conflict appears in the Conflict Alert List on the R-side display.

Conflict Probe uses momentary speed and heading, along with additional information (e.g., filed flight plan and cleared route variations), to predict losses of separation with look-ahead times as long as 20 minutes for high-likelihood conflicts. When a loss of separation between aircraft is predicted, a red number appears in the Aircraft List and the relevant data blocks within the Graphical Plan Display on the D-side. When a loss of separation between aircraft conformance boundaries is predicted, a yellow number appears. The value of the numbers indicates the total number of conflicts predicted. No notification is provided on the R-side.

The strategic nature of Conflict Probe provides important functionality for controllers. The long look-ahead time can aid in detecting potential conflicts earlier. Through trial-planning tools, it can also help controllers avoid actions that would result in future conflicts. At present, this functionality is available only on the D-side, which limits its usefulness. When controllers are working as a team, the R-side controller does not get the full benefit to situation awareness provided by Conflict Probe and cannot assist the D-side controller with trial planning. A controller working a sector alone, particularly in high workload situations, is likely to view the D-side minimally or not at all.

Considering the limitations of the current implementation, it is desirable to provide Conflict Probe functionality to both R-side and D-side controllers. However, controllers generally believe that the current Conflict Probe algorithm produces an unacceptably large number of false or nuisance alerts.

A concern therefore exists that introducing the current Conflict Probe to the R-side would create excessive distractions, consequently reducing situation awareness and overall performance.

Several changes to the algorithm have been proposed and are currently being tested, with the hope that Conflict Probe can be substantially improved. To properly evaluate these changes, it is critical to answer three questions:

1. **Performance Metric(s):** Of various possible metrics that one could use to describe automation performance (e.g. error rate, d', etc.), which metric (or metrics) most appropriately characterizes the automation performance in an operationally meaningful way (e.g., a way that will predict joint human-automation performance)?

2. **Performance Criterion:** What is the "crossover point" of the metric(s); that is, the automation performance threshold below which system performance will suffer?

3. **Current Performance:** What is the performance level of the proposed Conflict Probe algorithm(s) on the performance metric(s) chosen under Question 1?

A fair amount of work in the human factors literature addresses the effects of automation on human and overall system performance. Therefore, we conducted a literature review to determine whether Questions 1 and 2 have been definitively answered in a way that is clearly applicable to Conflict Probe. We then conducted further analysis on the results of the engineering studies of Crowell et al. (2011) and Crowell and Young (2012) to determine the Conflict Probe's performance on potentially useful metrics that had not been computed in the two papers.

2. HUMAN FACTORS LITERATURE REVIEW AND META-ANALYSIS

2.1 Background

Wickens and Dixon (2007) conducted an informal meta-analysis of human factors studies, with a goal of describing the relationship between automation performance and overall performance. As in the current analysis, they attempted to establish whether there was an automation performance threshold, below which operator performance was reduced to a level worse than the baseline (no-automation) condition. Wickens and Dixon (2007) had four criteria for studies to be included in their analysis:

1. The task environment made all of the necessary raw data available to the human user. The automation did not have access to any information that was unavailable to the user, and the presence of automation did not limit users' access to raw data.

2. The experiment included a baseline experimental condition that presented raw data to users without any automation aid, and at least one condition with an additional automation aid that was not perfectly reliable (i.e., % correct < 100). Many studies also included conditions with perfectly reliable automation, but these were not considered in the analysis.

3. Automation output was reported in sufficient detail to calculate the reliability of the automation (i.e., percent correct; see Formula (1) in Section 2.2.2. Ideally, studies' methodology reporting would include all four cells of the signal detection matrix: valid alerts (hits, H), valid non-alerts (correct rejections, CR), invalid alerts (false alarms, FA), and invalid non-alerts (misses, M).

4. Users were exposed to automation errors through training or, at a minimum, were informed that the automation was not perfectly reliable. This ensured that users' expectations were not grossly mismatched with actual automation performance.

Sixteen studies met these criteria. Many of the studies examined multiple reliability levels, other independent variables (e.g., workload), and multiple dependent variables (e.g., error rate and response time), yielding a total of 35 comparisons between baseline results and automation results. The majority of comparisons were on the automation-aided task, but some studies examined multi-task environments in which the effect of automation on concurrent tasks could also be measured and compared against baseline. For each of the 35 baseline-automation comparisons, Wickens and Dixon assigned an impact score reflecting whether a statistically significant difference ($p < 0.05$) had been found between the two conditions:

-1	=	Automation significantly worse than baseline
0	=	No significant difference between baseline and automation
1	=	Automation significantly better than baseline
-2 (or 2)	=	For studies with multiple automation reliability levels: Automation significantly worse (or better) than another automation level that in turn was significantly worse (or better) than baseline, e.g., if performance for the 0.6 condition was lower than performance for the 0.8 condition, which in turn was lower than baseline performance, the 0.8 condition would receive an impact score of -1 and the 0.6 condition would be assigned a -2 score. (*Note: this applied to a small number of data points.*)

A linear regression was performed on this data with reliability of the automation serving as the predictor variable and baseline-automation impact score serving as the predicted variable. The analysis exhibited a statistically significant positive relationship, with greater automation reliability associated with better system performance as compared to baseline. The regression line crossed the 0 impact mark at a reliability level of 0.70, with a 95% confidence interval from 0.63 to 0.77. In other words, the analysis suggested a 0.70 automation reliability threshold, with automation reliability less than this value being more likely to decrease performance than increase it, relative to baseline.

2.2 Method

The present analysis adopted much of the Wickens and Dixon (2007) framework, using the same four inclusion criteria. However, our methodology departed from theirs in several important ways: composition of studies, automation metrics, and measure of relative performance.

2.2.1 Composition of Studies

The present analysis included three studies that were not included by Wickens and Dixon (2007), two of which were published since 2007. There were also several studies that were used by Wickens and Dixon but not included here. Some of the excluded studies did not actually have baseline conditions that were sufficiently comparable to the automation conditions (Molloy & Parasuraman, 1996; Parasuraman, Molloy, & Singh, 1993; Wickens, Dixon, & Johnson, 2005; Yeh & Wickens, 2001). Other studies had baseline conditions but did not directly compare baseline and specific automation

conditions. Rather, they compared baseline to the average performance on a set of multiple automation conditions with different reliability levels, making it impossible to determine relative performance for any *specific* reliability level (Ben-Yaacov, Maltz, & Shinar, 2002; Dingus et al. 1997; Maltz & Shinar, 2003). Another study (Maltz & Meyer, 2001) performed the relevant comparisons, but did not provide sufficient detail to compute effect sizes. Of the 16 studies used by Wickens and Dixon, nine were considered appropriate for inclusion in the present analysis. Of those studies retained, several had individual data points that we excluded because of incompatible conditions or insufficient detail reported (Davison & Wickens, 2001; Dixon, Wickens, & McCarley, 2007; Skitka, Mosier, & Burdick, 1999). Several studies examined multiple dependent variables or presentation formats for the same automation level. Though Wickens and Dixon chose to average over these in some cases, we include every comparison as a separate data point. The full set of data in the present analysis included 34 baseline-automation comparisons from 12 studies.

2.2.2 Automation Performance Metrics

Wickens and Dixon (2007) characterized automation performance using a single metric: reliability, or the proportion of all opportunities that were responded to correctly by the system. Mathematically, this can be expressed as follows:

(1)

$$Reliability = \frac{H + CR}{H + CR + M + FA}$$

where H = total number of hits, CR = correct rejections, M = misses, and FA = false alarms. In addition to reliability, the present analysis explored the relationship between automation and overall performance with three other metrics describing automation performance, hit rate (HR), false alarm rate (FAR), and Positive Predictive Value (PPV). HR and FAR are standard signal detection theory (SDT) metrics and the formulas are as follows:

(2)

$$HR = \frac{H}{H + M}$$

(3)

$$FAR = \frac{FA}{FA + CR}$$

PPV, though not one of the usual SDT metrics, is commonly reported in studies that assess the predictive capability of an automated system or a human. It is defined as follows:

(4)

$$PPV = \frac{H}{H + FA}$$

4

This formula answers the question: "What percentage of the alerts that are generated are true?" or, worded alternatively, "How many false alarms are there for every hit?"

2.2.3 Measure of Relative Performance

Wickens and Dixon (2007) performed a linear regression on impact scores of statistical significance that, as described earlier, was a categorical variable with only five possible values. In contrast to this approach, we performed a traditional meta-analysis on effect size, a measure of the difference between conditions relative to the total variability in the data. There is some similarity between the impact score and effect size measures, as effect size is essentially what is being tested in a statistical significance test. Beyond this similarity, an analysis on effect size has two principal advantages. As a continuous variable, it is not susceptible to the potential distortions caused by sharply separating values that fall on different sides of the arbitrary $p < 0.05$ cutoff. Also, effect size is independent of the total number of data points collected (N), whereas statistical significance can be greatly affected by N. A small difference between conditions can be statistically (though perhaps not practically) significant with high N, and a meaningfully large difference can fall short of significance if low N results in insufficient power. Again, using effect size as the measure of interest avoids these distortions. The measure of effect size used in the present study was Cohen's d,

(5)

$$d = \frac{\overline{X}_A - \overline{X}_{NA}}{s}$$

where $\overline{X}_A$ = mean performance in the automation condition, $\overline{X}_{NA}$ = mean performance in the non-automated baseline condition, and s is the standard deviation of measured performance. Negative values of d indicate that the overall system performed better in the baseline condition, in the absence of automation assistance, whereas positive values indicate better performance in the automated condition. Although absolute values of d can range from 0 to infinity, a general guideline is that values around 0.2 are small, 0.5 are moderate, and 0.8 are large (Cohen, 1977). Cohen's d was not directly reported in any of the studies, but it could be derived from reported means and standard deviations, as above, or from t or F statistics.

2.3 Results

Table 1 contains all of the data points for the present analysis of the human factors literature. The authors and publication date of the study are listed in the first column (see References section for full citations). The reliability, HR, FAR, and PPV of the automation condition(s) in the study are in columns 2 through 5. The effect size of system performance in the automation condition compared to the non-automated baseline is in column 6. The task domain and participant population are in columns 7 and 8. The last column contains an indication of whether the measured task was considered primary or non-primary (for multi-task environments) or was the only task performed. Note that there can be multiple entries for a particular level of automation performance in a single study when there were multiple dependent variables (e.g., error rate and response time) or automation display formats (e.g., text vs. spatial cues).

Table 1. Summary of Studies, Automation Performance Levels, and Effects Compared to Baseline

Study	Automation Performance				Cohen's *d*	Task Domain	Participant Population	Task Priority
	Reliability	Hit Rate	False Alarm Rate	PPV				
Davison & Wickens, 2001	70%	70%	0%	100%	0.00[e]	Helicopter flight	Pilots	Non-primary
Davison & Wickens, 2001	70%	70%	0%	100%	0.97	Helicopter flight	Pilots	Non-primary
Dixon & Wickens, 2006	64%	94%	U	67%	-0.78 (S)	UAS operation	Pilots & students	Non-primary
Dixon & Wickens, 2006	64%	67%	U	94%	-0.43	UAS operation	Pilots & students	Non-primary
Dixon & Wickens, 2006	64%	94%	U	67%	-0.42	UAS operation	Pilots & students	Non-Primary
Dixon & Wickens, 2006	64%	67%	U	94%	-0.68 (S)	UAS operation	Pilots & students	Non-primary
Dixon & Wickens, 2006	80%	89%	U	89%	-0.33	UAS operation	Pilots & students	Non-primary
Dixon & Wickens, 2006	80%	89%	U	89%	-0.02	UAS operation	Pilots & students	Non-primary
Dixon et al., 2007	60%	100%	80%	56%	-2.23 (S)	Aircraft flight	Students	Non-primary
Dixon et al., 2007	60%	20%	0%	100%	-1.01	Aircraft flight	Students	Non-primary
Dixon et al., 2007	60%	100%	80%	56%	-2.37 (S)	Aircraft flight	Students	Non-primary
Dixon et al., 2007	60%	20%	0%	100%	0.37[e]	Aircraft flight	Students	Non-primary
Galster et al., 2001	67%	U	U	U	3.23 (S)	Artificial visual search	Unknown	Only
Galster et al., 2001	67%	U	U	U	2.89 (S)	Artificial visual search	Unknown	Only
Hitchcock et al., 2003	95%	40%	3%	40%	0.50 (S)	Artificial detection	Students	Only
Hitchcock et al., 2003	95%	40%	3%	40%	0.43 (S)	Artificial detection	Students	Only
Hitchcock et al., 2003	98%	80%	1%	80%	0.89 (S)	Artificial detection	Students	Only
Hitchcock et al., 2003	98%	80%	1%	80%	0.72 (S)	Artificial detection	Students	Only
Lehto et al., 2000	71%	100%	47%	56%	1.07 (S)	Automobile driving	Experienced drivers	Only

(table continues)

	Automation Performance							
Study	**Reliability**	**Hit Rate**	**False Alarm Rate**	**PPV**	**Cohen's *d***	**Task Domain**	**Participant Population**	**Task Priority**
Lehto et al., 2000	96%	89%	0%	100%	1.69 (S)	Automobile driving	Experienced drivers	Only
Rovira & Parasuraman, 2010	67%	67%	0%	100%	0.07	Air traffic control	Certified controllers	Primary
Rovira & Parasuraman, 2010	75%	100%	U	75%	0.49 (S)	Air traffic control	Certified controllers	Primary
Skitka et al., 1999	88%	94%	U	94%	0.00[e]	Aircraft flight	Students	Primary
St. John & Manes, 2002	75%	75%	25%	23%	1.84 (S)	Artificial visual search	Students	Only
St. John & Manes, 2002	75%	75%	25%	23%	1.02 (S)	Artificial visual search	Students	Only
St. John & Manes, 2002	95%	95%	5%	66%	5.78 (S)	Artificial visual search	Students	Only
St. John & Manes, 2002	95%	95%	5%	66%	6.46 (S)	Artificial visual search	Students	Only
Wang et al., 2009	75%	100%	50%	67%	0.75	Combat identification	Students	Only
Wang et al., 2009	88%	100%	25%	80%	1.59 (S)	Combat identification	Students	Only
Wiegmann et al., 2006	78%	90%	25%	47%	0.61	Baggage screening	Non-screener volunteers	Only
Wiegmann et al., 2006	78%	90%	25%	47%	1.23 (S)	Baggage screening	Non-screener volunteers	Only
Wiegmann et al., 2006	78%	90%	25%	47%	2.46[e] (S)	Baggage screening	Non-screener volunteers	Only
Xu et al., 2007	83%	U	U	U	1.06	Aircraft conflict detection	Student pilots	Only
Xu et al., 2007	83%	U	U	U	0.13	Aircraft conflict detection	Student pilots	Only

Note. PPV = Positive Predictive Value, U = Unknown (see explanation following table), UAS = Unmanned Aircraft Systems. (S) in "Cohen's d" column indicates the effect was statistically significant ($p <= 0.05$), given the sample size used in the study. An "e" in "Cohen's *d*" column indicates the effect size was estimated from incomplete data reporting. Adjusting or removing these values has no meaningful impact on any subsequent analyses.

For several of the studies, authors stated the overall reliability of the automation but did not provide sufficient detail to calculate some of the other performance metrics. This was particularly the case with false alarm rates. Many studies had false alarm events but no explicitly defined correct rejection events, so the denominator for the false alarm rate computation was ambiguous. A "U" entry in Table 1 indicates that the value on the relevant automation metric is unknown. Figure 1 contains four scatterplots that illustrate the association between metrics of automation performance and the baseline-automation effect size for each of the relevant data points.
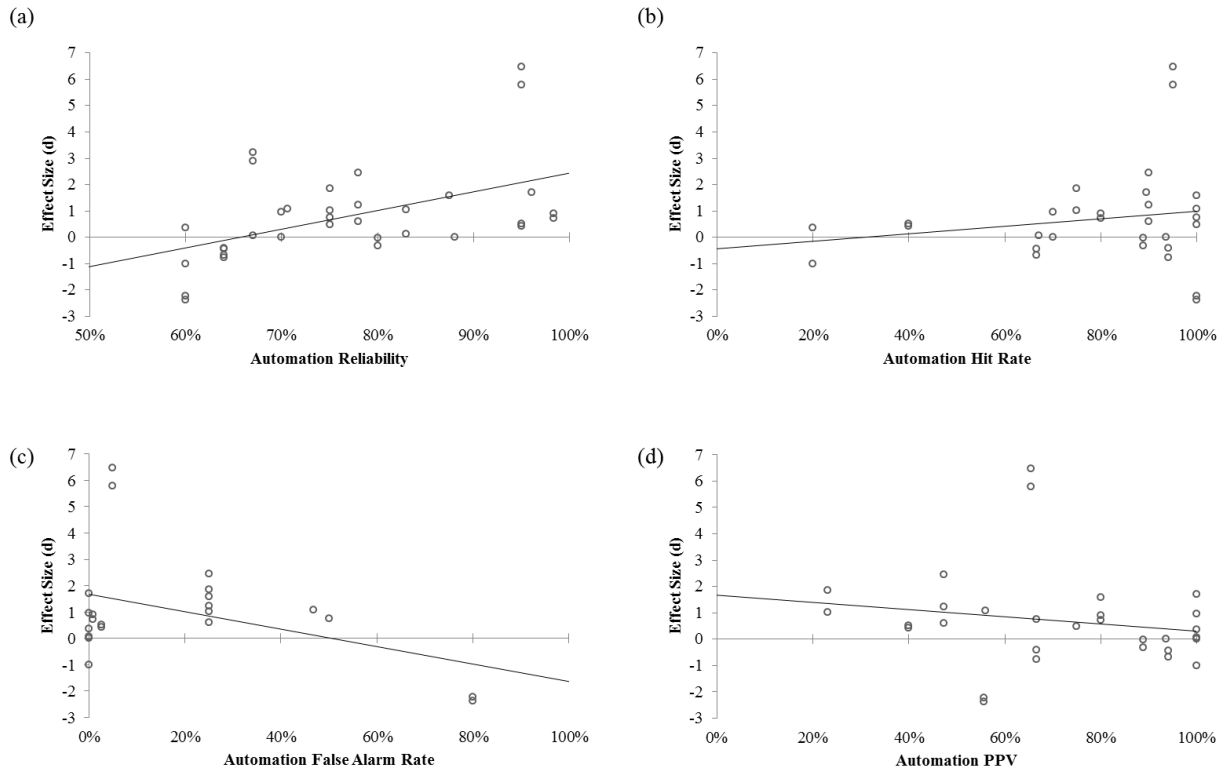
Figure 1. Scatterplots of automation performance and associated effect size relative to baseline: (a) reliability, (b) hit rate, (c) false alarm rate, and (d) positive predictive value.

2.3.1 Reliability

There was a statistically reliable positive relationship between automation reliability and effect size, $r(32) = 0.49$, $R^2 = 0.24$, $p = 0.003$. The regression equation expressing this relationship between reliability (*Rel*) and effect size (*d*) was $d = 7.11*Rel - 4.64$. The crossover point of effect size $= 0$ was at reliability $= 0.65$. To explore the noise in the data and the influence of outliers, a bootstrapping simulation was performed. In this analysis, the set of 34 data points was sampled with replacement one million times to generate one million 34-point samples, each with a corresponding regression function. The 95% confidence interval of the correlation coefficient ranged from 0.18 to 0.71. The interval for the crossover point ranged from 0.39 to 0.72.

Regarding outliers, two data points in particular had a large amount of influence on the regression line, as measured by Cook's D. These were the two data points from the St. John and Manes (2002) study with exceptionally large effect sizes. When they were removed and regression was performed on the trimmed data set, the correlation was reduced but remained reliable, $r(30) = 0.36$, $R^2 = 0.13$, $p = 0.046$, with a crossover point of 0.62.

2.3.2 Hit and False Alarm Rates

There was no statistically reliable relationship between automation HR and effect size, $r(28) = 0.18$, $R^2 = 0.03$, $p = 0.35$. Removing high-influence data points did not have any qualitative effect.

A marginally statistically reliable negative relationship existed between automation false alarm rate and effect size, $r(20) = -0.41$, $R^2 = 0.17$, $p = 0.06$. The regression equation expressing this relationship between false alarm rate (FAR) and effect size (d) was $d = -3.32*FAR + 1.68$. The crossover point of effect size = 0 was at false alarm rate = 0.51. However, this result was highly influenced by four data points: the St. John and Manes (2002) study with large positive effect sizes and the Dixon, Wickens, and McCarley (2007) study with exceptionally high automation false alarm rates and negative effect sizes. When they were removed and regression was performed on the trimmed data set, the trend was inverted, with a marginally reliable positive relationship between false alarm rate and effect size, $r(16) = 0.42$, $R^2 = 0.18$, $p = 0.08$ (see Figure 2).
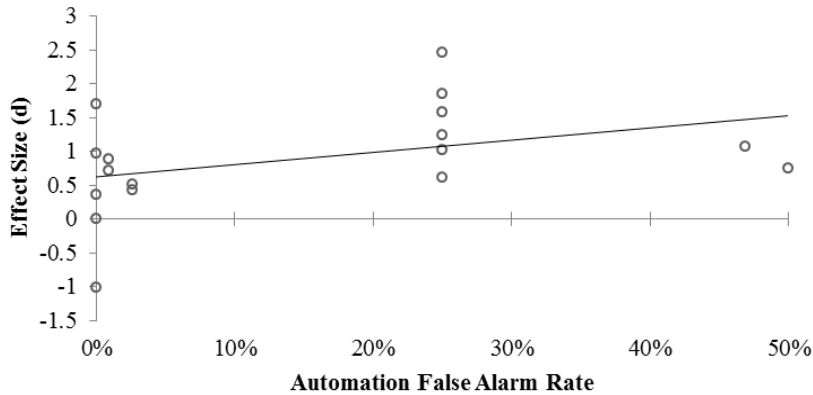

Figure 2. Scatterplots of automation false alarm rate and associated effect size relative to baseline, with four outliers from original data set removed.

2.3.3 Positive Predictive Value

There was no statistically reliable relationship between PPV and effect size, $r(28) = 0.18$, $R^2 = 0.03$, $p = 0.34$. Removing high-influence data points did not have any qualitative effect.

2.4 Discussion

Like Wickens and Dixon (2007), the present analysis indicated a positive relationship between (a) automation reliability and (b) relative performance with automation as compared to performance in a baseline condition without it. The 0.65 crossover point observed here is very close to the 0.70 value presented by those authors. However, several main factors suggest that these findings should be interpreted with caution.

Despite being a fairly comprehensive survey of the relevant literature, the Wickens and Dixon (2007) analysis and the present analysis include a rather small number of unique studies. Furthermore, few of these studies exhibit any negative effect of automation at all. Of the 16 studies analyzed by Wickens and Dixon, only 5 studies had conditions or dependent variables that showed any statistically significant cost with automation, three of which were removed from the present analysis because they lacked a direct comparison to the baseline condition. Of the 12 studies that we analyzed, only 2 studies, producing nine data points in this analysis, had negative automation-baseline effect sizes. These studies were produced by the same research group using a very similar task domain. If the analysis were restricted to the remaining 10 studies, it would be impossible to interpolate any crossover point. Finally, these studies include a wide variety of tasks and participant populations, as well as

variety in the number of concurrent tasks to be performed and the priority placed on those tasks. It is possible that the effect of automation reliability on performance differs systematically according to the value of these dimensions.

3. ANALYSIS OF CONFLICT PROBE ENGINEERING STUDY RESULTS

3.1 Background

The Concept Analysis Branch (ANG-C41) has conducted a series of engineering studies on how the prediction performance of the ERAM Conflict Probe can be further improved with algorithmic enhancements and parameter changes. Crowell et al. (2011) conducted an analysis of the ability of the Conflict Probe subsystem of ERAM to detect aircraft-to-aircraft conflicts under various parameter and algorithmic settings. Traffic sets were created from actual recorded data in Washington Air Route Traffic Control Center (ARTCC, ZDC) in which the tracks of aircraft were time-shifted to create a sufficient number of conflicts to conduct the analysis. The raw traffic data would be expected to contain few conflicts, because with real traffic, most conflicts are resolved before they occur. The traffic scenarios were run through a fast-time simulation to determine what conflicts would have happened and whether the automation would have detected a conflict. Their results were reported in terms of the traditional signal detection categories—that is, the numbers and rates of hits, misses, false alarms and correct rejections made by the automation. All pairs of aircraft that had "encounters" were included in the analysis, with an encounter being defined as a pair of aircraft having, at any time during the simulation run, a *simultaneous* lateral separation of less than 25 nmi and vertical separation of less than 4,000 ft. These boundaries were set to permit a reasonably sized set of non-conflict aircraft pairs. A total of about 17,000 encounters were included in the analyses by Crowell et al.

The results were reported first in traditional crisp terms—every event either was or was not a conflict (a conflict being a case where lateral separation was less than 5 nmi and vertical separation was simultaneously less than 1,000 ft) and the automation response was defined as either predicting or not predicting a conflict. Because very few conflicts were completely missed by the automation, a late response (i.e., a conflict that was first predicted less than a specified time before its occurrence) was defined as being a non-response. Such *late alerts* were used as a proxy for misses, which were close to nonexistent. The late alert count as defined by Crowell et al. (2011) included both late alerts and a small number of actual missed detections.

Additionally, Crowell et al. (2011) reported fuzzy signal detection analysis results, in which a given encounter could fall into more than one of the traditional signal detection categories. The degree to which an event was a conflict (signal in SDT terms) could fall between zero and one based on the separation distance between the aircraft. For example, if a pair of aircraft had zero vertical separation and a lateral separation just beyond the 5-nmi cutoff, it would be classed partly as a non-conflict because it did not meet the binary cutoff for defining a conflict, but partly a conflict because it would likely be watched as an event of potential interest by many controllers. The degree to which an encounter was predicted by ERAM to be a conflict was permitted to vary anywhere between zero and one, with the degree of "conflict predicted" (response in SDT terms) defined based on predicted separation and warning time. For both the crisp and the fuzzy analyses, Crowell et al. computed the late alert rate (LAR) as follows:

(6)

$$LAR = \frac{LA}{H + LA}$$

Crowell et al. (2011) treated LAR as the miss rate. They also reported the FAR (see Section 2.2.2) for each simulation run.

Crowell and Young (2012) conducted a similar analysis for time-shifted traffic recordings from Chicago ARTCC (ZAU), which generated about 11,000 encounters. The purpose of this analysis, unlike that of Crowell et al. (2011) which compared multiple parametric settings and prototype enhancements, was to test a set of adherence parameters that define whether a flight is considered as being on its route. These parameters result in the removal from analysis of encounters where a flight was significantly off its route. Had these flights been included, the primary effect would have been an FAR that was spuriously high from the operational perspective, because invalid alerts would be occurring for flights that were off their assigned route. Multiple categories were used for classifying the result of each predicted or actual conflict, going beyond the usual SDT categories, for reasons pertinent to their analysis. However, each event was also classed at a high level as a hit—termed Valid Alert (VA)—miss, or false alarm, permitting the generation of some of the traditional SDT metrics. Crowell and Young used different criteria than Crowell et al. for including aircraft pairs in the analysis: an encounter was defined as a simultaneous lateral separation of less than 30 nmi and vertical separation of less than 5,000 ft. CR counts were not included in the original memo but were later provided (Crowell & Willems, personal communication, 2012), which made it possible to complete the computations of SDT metrics and the other metrics used for the present study. Crowell and Young did not compute fuzzy SDT metrics.

3.2 Method

For the present analysis, the counts and rates found by Crowell et al. (2011) and Crowell and Young (2012) were converted into the reliability metrics that were reported in, or computed from, the studies included in the meta-analysis described in Section 2. Specifically, the metrics computed were reliability (best and worst case), PPV, hit rate, and false alarm rate. Reliability uses the formula defined by Wickens and Dixon (2007), shown in Section 2.2.2. This formula reflects the proportion of the total events (encounters) for which the automation made a correct judgment.

It has been suggested within ANG-E25 that the criteria for including encounters in the analysis set of Crowell et al. (2011), while useful for their study's purposes, were too lax for the purpose of evaluating accuracy from the operational perspective, because encounters where the aircraft are near the outer limits of their criteria (e.g., 24 miles apart laterally and 3,000 ft vertically) are generally not situations that controllers would be even slightly concerned with. Although they were considered CRs by the analysis, the argument can be made that they are not really events of interest. Therefore, the Wickens and Dixon (2007) formula might be considered a *best case* for reliability, in which the automation is given "credit" for all events that could even potentially be considered of interest. For the purpose of reporting the present results, the above formula for R will therefore be labeled R_{best}. For purposes of clarity, the reliability formula given in Section 2.2.2 is reproduced with its alternative name.

(7)

$$R(best) = \frac{H + CR}{H + CR + M + FA}$$

The other extreme in this type of reliability analysis is to assign no credit for any of the CRs. In some studies and real-world applications, no clear definition exists of what would constitute a CR. Although ATC conflict detection is not an example of such an application, some of the domains

studied in the meta-analysis described in Section 1 may be considered as such. Therefore, to enable a fair comparison across studies and applications, a *worst case* reliability was defined as follows:

(8)

$$R(worst) = \frac{H}{H + M + FA}$$

Hit rate was not directly reported by Crowell et al. (2011), who focused on error rates rather than correctness rates, because a goal of their analysis was to determine how to best reduce the main types of errors—misses and false alarms. However, it is easily calculated by subtracting the miss rate from 1. As mentioned in Section 3.1, Crowell et al. used late alerts as a proxy for misses, summing them together with the small number of actual misses to derive a total miss count. For this reason, the hit rate for the present analysis was defined as follows:

(9)

$$HR = 1 - LAR$$

The formulas used for FAR and PPV in this analysis are identical to those given in Section 2.2.2; therefore, they are not reproduced here.

3.3 Results

Because multiple combinations of parameter settings and prototype enhancements were studied in the Crowell et al. (2011) ZDC analysis, they reported results for 30 experimental runs, in addition to two baseline runs—so termed because they did not contain the parameter settings or the prototype enhancements being tested. The first baseline used settings equivalent to those in the Conflict Probe implemented in the present operational version of ERAM. The other baseline contained a set of trajectory modeling enhancements that were also incorporated into the experimental runs. This second baseline was included in the original study to provide a more valid comparison with the experimental runs. The runs varied in terms of automation performance.

Because the goals of Crowell et al. and of the present study were to assess the optimal level of automation performance attainable by the Conflict Probe under the current implementation, only the highest performing run on each metric was selected for the present comparison. Crowell and Young (2012) were interested in testing an entire set of parameters, rather than various combinations of parameters as in Crowell et al., and therefore reported only two sets of SDT results; one with and one without the adherence parameters.

The results for the selected ZDC and ZAU runs on the various metrics of interest are presented in the following subsections.

3.3.1 Reliability

Table 2 shows the reliability values, as computed according to both the best-case and worst-case computations. For the ZDC analysis, which had multiple experimental runs (30) and baseline runs (2), the values reported in the table are the maximum values achieved in any of the 30 experimental runs and in either of the baseline runs.

Table 2. Experimental and Baseline Reliability Values (for ZDC, Maximum Experimental Values and Maximum Baseline Values)

ARTCC	Crisp or fuzzy metrics	R_{best} Experimental	R_{best} Baseline	R_{worst} Experimental	R_{worst} Baseline
ZDC	Crisp	0.964	0.918	0.192	0.104
ZDC	Fuzzy	0.970	0.943	0.330	0.212
ZAU	Crisp	0.948	0.876	0.244	0.119

Note. ARTCC = Air Route Traffic Control Center, ZDC = Washington ARTCC, ZAU = Chicago ARTCC.

The automation exhibited very high values on the best-case reliability, where all CRs were included in the analysis. The best achievable R_{best} for each analysis was 0.964 (ZDC crisp), 0.970 (ZDC fuzzy), and 0.948 (ZAU). These values are well above the reliability cutoffs determined in Wickens and Dixon (2007) and the present meta-analysis and, in fact, are above the upper bound of the confidence intervals.

For the worst-case reliability, the highest values achieved were 0.192 (ZDC crisp), 0.330 (ZDC fuzzy), and 0.244 (ZAU). Taken at face value, these figures are well below the cutoff reliability level, and the confidence interval lower bound, for automation benefit according to Wickens and Dixon (2007) and the present meta-analysis.

However, when comparing these observed Conflict Probe reliability values to the cutoff values from previous analyses, it should be noted that the studies reviewed for the meta-analyses described in Section 2 did not have a consistent definition of what constituted an event. Some did not include any correct rejections in their reporting, whereas others had a prevalence of correct rejections as high as 95% of total events. The reliability levels reported, and therefore the cutoff values computed, represent a mix of results from both R_{best} and R_{worst} formulae.

3.3.2 Hit and False Alarm Rates

Table 3 presents the HR and FAR values. As with reliability, the maximum performance—the higher/highest HR and lower/lowest FAR—is given for experimental runs and baseline runs for each analysis.

Table 3. Experimental and Baseline HR and FAR Values (for ZDC, Maximum Performance Experimental Values and Maximum Performance Baseline Values)

ARTCC	Crisp or fuzzy metrics	HR Experimental	HR Baseline	FAR Experimental	FAR Baseline
ZDC	Crisp	0.910	0.916	0.035	0.082
ZDC	Fuzzy	0.933	0.931	0.028	0.057
ZAU	Crisp	0.989	0.939	0.053	0.126

Note. ARTCC = Air Route Traffic Control Center, HR = Hit Rate, FAR = False Alarm Rate, ZDC = Washington ARTCC, ZAU = Chicago ARTCC.

Like the related metric of reliability, Hit rates for the ERAM Conflict Probe were also very high: 0.916 for crisp in ZDC, 0.933 for fuzzy in ZDC, and 0.989 in ZAU. As with R_{best}, these values seem to be much higher than would be necessary to see an automation benefit according to our meta-analysis of the automation literature. However, it must be reiterated that a statistically reliable relationship was

not found in the present study between HR and automation benefit as measured by Cohen's *d*, and that some of the reviewed studies featuring HR in the > 90% range showed a performance decrement due to automation. It is not possible to make definitive conclusions about whether the Conflict Probe's HR is expected to be operationally large enough, because it has not been conclusively shown that high HR—without considering other elements of the SDT matrix—is associated with performance benefits.

The Conflict Probe exhibited low FARs: 0.035 for crisp in ZDC, 0.028 for fuzzy in ZDC, and 0.053 for ZAU. The meta-analysis did not find a stable statistical relationship between FAR (in and of itself) and degree of automation benefit. However, it is encouraging (see Figure 2 and the FAR portion of Figure 1) that moderate to large automation benefits were seen in most of the reviewed studies in which automation had an FAR near zero, as the Conflict Probe had when all encounters meeting the specified separation criteria were included in the analysis. Only two exceptions were found in the reviewed studies to the general rule of low FAR benefiting performance (out of the nine data points in which FAR was less than 5%), with one effect size of zero and one negative effect size. Taken at face value, the ERAM automation's FAR appears to be operationally acceptable. However, as noted in Section 3.2, the set of encounters used to compute this FAR may be crediting the automation with some CRs that are not of potential interest to controllers. Non-alerts by the automation on these events, therefore, add no informational value. If the FAR calculation were restricted to operationally relevant CRs only, the FAR would be higher than observed here, potentially much higher.

3.3.3 Positive Predictive Value

PPV results calculated from the data in Crowell et al. (2011) and Crowell and Young (2012) are given in Table 4, in the same format as Tables 2 and 3.

Table 4. Experimental and Baseline PPVs (for ZDC, Maximum Experimental Values and Maximum Baseline Values)

ARTCC	Crisp or fuzzy metrics	Experimental	Baseline
ZDC	Crisp	0.198	0.105
ZDC	Fuzzy	0.348	0.215
ZAU	Crisp	0.245	0.120

Note. ARTCC = Air Route Traffic Control Center, ZDC = Washington ARTCC, ZAU = Chicago ARTCC.

The results present an encouraging picture of the computational enhancements assessed in the original studies, in that the best achievable PPVs were approximately double for the experimental run (or for ZDC, the best performing experimental run) in each analysis as compared to the corresponding baseline run (or better performing baseline run in ZDC).

The PPV metric can be communicated to controllers in a way that is highly intuitive (ratio of false alarms to true alerts), and the results of the Crowell et al. (2011) and Crowell and Young (2012) studies can thus be presented to controllers in a positive light. For example, using the ZDC crisp metric, it can be accurately stated that the algorithmic and parametric improvements can reduce the ratio of FAs to hits from 9:1 to 4:1. However, given the lack of a statistically reliable relationship between PPV and automation-related performance enhancement, further study is still needed to ascertain whether the observed PPVs are expected to be associated with performance improvements.

3.4 Discussion

The best-case reliability values found in the present analysis of the engineering studies appear to be acceptable, and likely to be associated with performance benefits, according to the present meta-analysis of experimental automation studies. However, this is not so for the worst-case values. When interpreting the worst-case finding, it must be recognized that for some experimental and operational tasks, it is difficult to define what constitutes a correct rejection. Clearly ATC is one of these tasks; therefore, depending on what situations are included in a reliability analysis, different reliability numbers might be derived.

For the present application of ATC, it is necessary to obtain further operational input to better define what encounters—that is, what magnitude of separation between aircraft—should be included in the analysis to generate the most operationally meaningful definition of CRs. This more operationally sound definition of CRs will also affect the FAR that is computed because the CR count is part of the denominator for computing FAR. From the present analysis, it may be tentatively concluded that the Conflict Probe's current achievable FAR would be associated with performance benefit, but additional operational input for a better definition of what constitutes a CR will help to strengthen this assumption.

The data that we analyzed from the ANG-C41 engineering studies showed what could be considered operationally significant PPV improvements as result of the parametric and prototype changes studied. However, it is intuitive that controllers may still consider the PPVs found here to be low (i.e., high number of FAs relative to hits). The results of the meta-analysis, which showed no reliable relationship between PPV and automation's benefit to performance, do not permit conclusions about the operational appropriateness of the ERAM Conflict Probe's current performance on the PPV metric. Therefore, a more operationally oriented study is necessary to determine whether the prevalence of FAs in relation to hits is operationally acceptable. It might be that some of the types of events that are generating false alarms are situations that are neither conflicts nor near-conflicts but that for operational reasons could develop into conflicts, and the controller may thus wish to be alerted to them. The recommended research to directly assess this issue will be described in the next section.

4. GENERAL DISCUSSION

Like the Wickens and Dixon (2007) analysis, the meta-analysis of the human factors literature presented here suggests that even imperfect automation can have a positive effect on performance. In particular, automation reliability was a significant predictor of performance gain over baseline, with an observed cutoff of approximately 0.65. The best-case reliability of Conflict Probe is well in excess of this value, and in fact, with values of over 0.90, exceeds the upper bound of the confidence interval for the cutoff.

However, the exact value of 0.65 may not be an appropriate standard, for two principal reasons. First, this estimate was subject to a large amount of uncertainty, as indicated by the very broad confidence interval. Second, the studies used in the meta-analysis may not be sufficiently applicable to the topic of Conflict Probe. Only one of the 12 studies actually investigated ATC. Most of the studies measured performance either in a single-task environment or on a non-primary task, in contrast to the complex and dynamic ATC environment in which conflict detection has high priority. In addition, the majority used participants who were not experts in the task domain. These studies may have been more likely to exhibit performance improvements with imperfect automation because of a low level of baseline performance. In contrast, experienced air traffic controllers already perform at a high

level and, therefore, may be more likely to have that performance harmed by the intrusion of even 0.80 or 0.90 reliable automation. Similarly, HR, FAR, or PPV may have a systematic effect on human-automation performance in ATC, despite the lack of a relationship observed in the meta-analysis. For example, the St. John and Manes (2002) study—which had, by far, the greatest positive effect from automation and the largest influence in the meta-analysis regression—used a single, highly artificial task with novice participants.

Of the 12 studies in the meta-analysis, only 2 studies exhibited a negative effect of automation (Dixon & Wickens, 2006; Dixon et al., 2007), both in the lower reliability condition—as would be expected. One factor, not directly addressed in these two studies or in many of the experiments reviewed, which may affect the positive or negative impact of automation on performance, is users' level of subjective confidence or trust in the automation (Lee & See, 2004). If trust in the automation is properly calibrated; that is, the operator knows through training and/or experience how reliable the automation should be and under what conditions it should be more and less reliable (Masalonis & Parasuraman, 2003), performance should not be harmed as much as it otherwise would be by low-reliability automation, and should be helped even more by higher reliability automation than otherwise. Though participants were informed of automation unreliability in each of the studies in the meta-analysis, the level of training and experience were less than would be present in an operational setting. Appropriate training on the strengths and weaknesses of Conflict Probe may improve its acceptability and impact on overall performance, even if it exhibited lower levels of accuracy.

Among the metrics we examined, the PPV automation performance metric might well have the most promise for being operationally intuitive. Unfortunately, this is also the metric about which the present meta-analysis and Conflict Probe engineering study data analysis are least able to inform conclusions. PPV can be further explored in several ways. One is to investigate in detail the situations that cause the automation to generate FAs. The FlightGUI software developed by ANG-C41 will be useful in conducting initial exploration of this issue. Situations identified in this exploration could be further analyzed via a study wherein the situations of interest are presented to controllers, and they are asked to provide an assessment of the severity of the conflict, or the likelihood that they would wish to be informed about it. If a significant number of the encounters that are generating FAs and decreasing the PPV in the present study are of operational interest to the controllers, then it may be appropriate to conduct engineering analyses similar to the work of Crowell et al. (2011) and Crowell and Young (2012) using new definitions of what constitutes a conflict, a "partial" conflict, and a non-conflict. These new definitions might be derived based on modified algorithms for defining the fuzzy SDT categories into which various encounters fall. Another benefit of such a study would be to find the ideal spot between the best- and worst-case settings for defining CRs to derive the most meaningful reliability figures. More detailed descriptions of how such a study might be conducted are in preparation by ANG-E25 and the Spectrum Software Technology, Inc. (SST) team.

As mentioned previously, most of the studies reviewed in the meta-analysis (described in Section 2) were not in the ATC domain. The present analyses provide useful initial answers to the question of the Conflict Probe's current performance vis-à-vis a benchmark for what automation performance levels should be to benefit human-automation system performance. However, moderate- to high-fidelity simulation studies that explicitly vary reliability and/or the other performance metrics assessed here, in the ATC domain, will be necessary to make definitive conclusions regarding operational needs for Conflict Probe reliability. These issues will be considered in the planning of the third Separation Management HITL simulation—reliability and other performance metrics may, in fact, become a primary focus of this HITL.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Ben-Yaacov, A., Maltz, M., & Shinar, D. (2002). Effects of an in-vehicle collision avoidance warning system on short- and long-term driving performance. *Human Factors, 44*(2), 335–342.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Crowell, A., Fabian, A., Young, C. Musialek, B., & Paglione, M. (2011). *Evaluation of prototype enhancements to the En Route Automation Modernization's conflict probe* (NTIS No. DOT/FAA/TC-TN12/3). Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

Crowell, A., & Young, C. (2012). *Separation management modern procedures ZAU baseline scenario analysis report* [Memorandum]. Atlantic City International Airport, NJ: Federal Aviation Administration William J. Hughes Technical Center.

*Davison, H. J., & Wickens, C. D. (2001). Rotorcraft hazard cueing: The effects on attention and trust. *Proceedings of the 11th International Symposium on Aviation Psychology.* Columbus: The Ohio State University.

Dingus, T. A., McGehee, D. V., Manakkal, N., Jahns, S. K., Carney, C., & Hankey, J. M. (1997). Human factors field evaluation of automotive headway maintenance/collision warning devices. *Human Factors, 39*(2), 216–229.

*Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors, 48*(3), 474–486.

*Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors, 49*(4), 564–572.

*Galster, S. M., Bolia, R. S., Roe, M. M., & Parasuraman, R. (2001). Effects of automated cueing on decision implementation in a visual search task. In *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society*, 321–325. Santa Monica, CA: Human Factors and Ergonomics Society.

*Hitchcock, E. M., Warm, J. S., Matthews, G., Dember, W. N., Shear, P. K., Tripp, L. D., ... Parasuraman, R. (2003). Automation cueing modulates cerebral blood flow and vigilance in a simulated air traffic control task. *Theoretical Issues in Ergonomics Science, 4*(1), 89–112.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

*Lehto, M. R., Papastavrou, J. D., Ranney, T. A., & Simmons, L. A. (2000). An experimental comparison of conservative versus optimal collision avoidance warning system thresholds. *Safety Science, 36*(3), 185–209.

Maltz, M., & Meyer, J. (2001). 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.

Masalonis, A. J., & Parasuraman, R. (2003). Effects of situation-specific reliability on trust and usage of automated air traffic control decision aids. In *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*, 533–537. Santa Monica, CA: Human Factors and Ergonomics Society.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors, 38*(2), 311–322.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *The International Journal of Aviation Psychology, 3*(1), 1–23.

*Rovira, R., & Parasuraman, R. (2010). Transitioning to future air traffic management: Effects of imperfect automation on controller attention and performance. *Human Factors, 52*(3), 411–425.

*Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies, 51*(5), 991–1006.

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

*Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors, 51*(3), 281–291.

Wickens, C. D. & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science, 8*(3), 201–212.

Wickens, C. D., Dixon, S. R., & Johnson, N. R. (2005). *Task priorities and imperfect automation* (Technical Report AHFD-05-20/MAAD-05-6). Savoy: University of Illinois, Aviation Human Factors Division.

*Wiegmann, D., McCarley, J. S., Kramer, A. F., & Wickens, C. D. (2006). Age and automation interact to influence performance for a simulated luggage screening task. *Aviation, Space, and Environmental Medicine, 77*(8), 825–831.

*Xu, X., Wickens, C. D., & Rantanen, E. M. (2007). Effects of conflict alerting system reliability and task difficulty on pilots' conflict detection with cockpit display of traffic automation. *Ergonomics, 50*(1), 112–130.

Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors, 43*(3), 355–365.

Acronyms

ANG-C41	Concept Analysis Branch
ANG-E25	Human Factors Branch
ARTCC	Air Route Traffic Control Center
ATC	Air Traffic Control
COTS	Commercial-Off-The-Shelf
CR	Correct Rejections
D-side	Data-side
ERAM	En Route Automation Modernization
FA	False Alarms
FAA	Federal Aviation Administration
FAR	False Alarm Rate
H	Hits
HITL	Human-In-The Loop
HR	Hit Rate
LAR	Late Alert Rate
M	Misses
PPV	Positive Predictive Value
R-side	Radar-side
SDT	Signal Detection Theory
UAS	Unmanned Aircraft Systems
VA	Valid Alert
WJHTC	William J. Hughes Technical Center
ZAU	Chicago ARTCC
ZDC	Washington ARTCC

Author Biography

Jonathan R. Rein received his Ph.D. in Cognitive Psychology from the University of Texas at Austin. He has previously conducted performance evaluations of human-automation systems in the aviation security field. He has also conducted experiments on a variety of cognitive tasks including visual attention, search, and categorization.

Anthony J. Masalonis received his Ph.D. in Applied-Experimental Psychology from the Catholic University of America, specializing in Human Factors. His areas of research and development experience include the study of trust in automation for aircraft conflict detection and traffic flow management, the development of Fuzzy Signal Detection techniques for conflict detection and other Human Factors applications, and Air Traffic Management system procurement and Systems Engineering.

James Messina received his Ph.D. in Psychology from the Fordham University, New York. He has conducted human performance and human-computer system evaluations across many commercial software products and in government defense sectors. He has held offices in the New Jersey chapter of the Human Factors Society, and he was a founding member and officer of the User Experience Professionals Association of New Jersey.

Ben Willems is an Engineering Research Psychologist with the Human Factors Team-Atlantic City. He joined the Federal Aviation Administration in 1998 and has conducted human factors experiments in the terminal and en route Air Traffic Control environments. Experiments that Mr. Willems has conducted have investigated concepts such as the effect of traffic load levels, controller involvement, automation, and data communications on air traffic controller behavior. Data collection capabilities used during these experiments range from subjective measures (such as questionnaires and controller workload ratings) to objective measures (such as eye-movement recordings and measurements of oxygenation of the brain). The data collected during these experiments have been used to drive the development of system requirements, assess potential benefits of system enhancements, and build human performance models of air traffic controllers.