

## SEPARATION MAINTENANCE IN HIGH-STRESS FREE FLIGHT USING A TIME-TO-CONTACT-BASED COCKPIT DISPLAY OF TRAFFIC INFORMATION

William Knecht and Peter Hancock  
University of Minnesota  
Minneapolis, Minnesota

We evaluated the navigational performance of commercial airline pilots in simulated free flight conditions using two cockpit displays of traffic information (CDTI). One display presented minimal essential navigation information. The other added color-based time-to-contact (Tc), minimum-range, and altitude intent information. New metrics of separation maintenance were developed. Results indicated performance gains for the CDTI that automatically calculates and displays conflict-alerting information. Recommendations are made for advanced avionics for use in free flight.

### INTRODUCTION

The Federal Aviation Administration (FAA) is beginning to allow limited point-to-point *free flight* by commercial aircraft. Free flight involves allowing en-route navigation decisions currently made by air traffic controllers to be made by pilots. It is postulated that free flight will be more efficient in terms of time and money (RTCA, 1995).

Part of the evaluation of the free flight concept involves assessment of a ground- or cockpit-based *conflict probe*. A conflict probe is a computer-based methodology capable of predicting difficulties in keeping aircraft properly separated.

In practice, separation errors are defined by the FAA as any failure to maintain 9.3 km (5 nm) lateral and 305-610 m (1000-2000 ft) vertical separation (altitude-dependent). This is the so-called cylindrical "protected zone" surrounding each aircraft.

Research into the safety of free flight has generated mixed results. Battiste, Johnson, Delzell, Holland, Belcher, and Jordan (1997) and Lozito, McGann, Mackintosh, and Cashion (1998) indicated that when pilots were given a CDTI with a conflict probe, long lookahead times (e.g. 8 minutes), plus a premaneuver solution-testing tool, they appeared capable of maintaining error-free separation under a variety of traffic situations.

In contrast, Scallen, Smith, and Hancock (1996), Smith, Lewin, and Hancock (1996) and Smith (1998) came to a somewhat different conclusion. Given a minimalistic CDTI (which was nonetheless more sophisticated than those which many aircraft currently use) they found separation failure rates as high as 67%.

Barhydt and Hansman (1997) reported error rates of approximately 25-58% using a variety of displays with and without conflict probe, and demonstrated statistical significance for both error rate and time-to-maneuver in favor of probe-based displays.

One factor obscuring comparison between these studies is scenario difficulty. "Safe" free flight is highly dependent on situational difficulty, coupled with the degree of technology and training brought to bear on that difficulty.

Scenario specifics like traffic density, the initial placement of aircraft, their relative headings and speeds, and the degree and type of altitude changes, combine in complex ways to make navigation more or less difficult. Furthermore, the presence of an onboard conflict probe *and* a pre-maneuver solution-testing tool seem to help matters greatly.

The current study was designed to determine if a conflict probe alone was capable of enabling pilots to navigate safely through high-stress scenarios, and to what relative degree would the probe influence performance in comparison to other factors such as situational difficulty and individual differences.

### METHOD

Using the glass-cockpit flight simulator in the Human Factors Research Laboratory at the University of Minnesota (described in Scallen et al., 1996) sixteen commercial airline pilots (1 female, 15 male, mean total flight hours 10,600, SD 4,400) used two versions of CDTI to help them navigate through five simulated en-route crossing conflict situations with the goal of maintaining safe separation, defined as 9.3 km (5 nm) lateral, 305 m (1000 ft) vertical, not altitude-dependent. A conflict probe was available in one of the two displays; otherwise circumstances were identical.

Information on intruder altitude, heading, and airspeed was available in a text data tag near each aircraft symbol. This was updated once per second to allow estimation of altitude changes.

The conflict alerting logic was deterministic, meaning it expressed the probability of conflict as either "yes" or "no". A deterministic probe was chosen on the assumption that future commercial aircraft will have cybernetically controlled trajectory capability, as well as access to highly accurate global positioning satellite information.

On the basis of prior research the probe's lookahead value was set at four minutes. Knecht and Hancock (1997) found that two minutes was sufficient for resolving most simple two-ship conflicts in the absence of distractor aircraft. Barhydt and Hansman (1997) also noted little additional benefit to error rates beyond two minute lookahead in single-conflict situations including three distractor aircraft.

In the advanced Tc display, probe-predicted non-conflicts were displayed onscreen as white symbols, while conflicts were colored orange. Additionally, yellow was used to indicate illusory conflicts--those due to aircraft appearing to conflict, but which would not, due to pilot intent to level off at a safe distance.

A within-subjects experimental design was used, with repeated factors of both display and scenario. Within-subjects designs are statistically sensitive, but can introduce presentation order effects, including fatigue, learning over time, and asymmetrical transfer effects (Poulton, 1982). These

effects were addressed by giving no indication to pilots that scenarios would be repeated, by giving them as much practice time as they cared to take before data collection, and by counterbalancing the treatment presentation order for both display and scenario.

The five scenarios were presented as either "nc-first"--by showing them with the standard no-color display first, followed by same-order repetition with the advanced Tc display--or else in "tc-first" order. Before each change of display, a retraining session was conducted. Half the participants received each display order. Within each half, the order of scenario presentation was counterbalanced.

"High stress" airspace was defined here by restriction of available maneuvers. The scenarios had no more than six intruders per five-minute scenario, but we placed intruder aircraft to block two, and sometimes three, of the basic ownship maneuver options per scenario.

During each trial, separation was assessed with three metrics, each of which described a different aspect of performance. *Red alerts* described the integer number of protected zone violations. This modeled the FAA's "zero error-tolerance" stand on safety.

A new minimum-range metric *rmin* described the single lateral (xy-plane) point-of-closest-approach during each scenario. This was defined as the normalized geometric separation distance defined by

$$r_{min} = \sqrt{\left(\frac{r_{xy}}{C_{xy}}\right)^2 + \left(\frac{r_z}{C_z}\right)^2} \quad (1)$$

where  $C_{xy}$  was the "critical distance", or size of the protected zone in the xy-plane,  $C_z$  was the critical distance in the vertical (z) dimension,  $r_{xy}$  was the lateral range (separation distance) and  $r_z$  the vertical range.

Finally, a new measure of airspace intrusion was constructed, which we termed *niti*, the *normalized intrusion-time integral*. *Niti* was designed to embody the notion that conflict severity is a function of both separation and time, i.e.

"how close multiplied by how long". It was defined as the time integral of the dimensionally partialled proportion of penetration into an intruder's protected zone, formulated by

$$niti = \int_{t=start}^{t=end} \left( \left( \frac{C_{xy} - r_{xy}}{C_{xy}} \right) * \left( \frac{C_z - r_z}{C_z} \right) \right) dt \quad (2)$$

where "start" was the red-alert starting time and "end" its ending time.

*Niti* gives a simple, linear estimation of protected-zone penetration. It factor-weights in favor of simultaneous deep penetrations in all three dimensions, and against superficial penetrations in any single dimension. Its instantaneous value ranges from 0-1.0, peaking when separation is simultaneously zero in x, y, and z, (direct impact).

Typical values for minor penetrations are small numbers on the order of .1-1.0 or so, depending on one's personal standard for "severity". High numbers (e.g. our highest observed value of 103.5) represent catastrophic separation failure, and generally correspond to small *rmins* coupled with long event durations on the order of 30-60 seconds or more.

Performance data for all three metrics were adjusted to discard patently minor errors (e.g. tiny altitude overshoots into a protected zone). These were defined as any with values of *niti* < 0.05.

## RESULTS

### General Observations

Based on the raw score results presented in Table 1, there appeared to be an advantage to having an onboard conflict-probe in short-lookahead, high-stress tactical free flight situations. Even though this probe did not calculate solutions, it did appear to achieve what was intended, which was simply to ease the job of detecting the conflicts themselves.

Variable	Red alerts		Niti		Rmin	
	nc	tc	nc	tc	nc	tc
Mean	<b>0.513</b>	<b>0.300</b>	<b>5.290</b>	<b>1.783</b>	<b>1.076</b>	<b>1.181</b>
Std. Dev.	0.811	0.560	13.534	5.481	0.462	0.414
Median	0.000	0.000	0.000	0.000	1.018	1.083
Skewness	1.856	1.736	5.273	4.823	0.277	0.304
Kurtosis	3.983	2.112	35.398	28.221	-0.033	0.167
p(Skew Normality)	<0.001	<0.001	<0.001	<0.001	0.156	0.134
p(Kurt Normality)	<0.001	<0.001	<0.001	<0.001	0.476	0.380
Effect Size $\hat{\Delta}$	<b>0.30</b>		<b>0.34</b>		<b>-0.24</b>	

**Table 1.** Summary statistics for distributions of raw scores for the standard, no-color display (nc) and the Tc conflict-probe display (tc).

Variable	Red alerts		Niti		Rmin	
	Raw	Transfmd	Raw	Transfmd	Raw	Transfmd
Mean	0.213	-0.252	3.507	-4.054	-0.105	-0.063
Std. Dev.	0.896	1.305	14.665	21.524	0.562	0.304
Median	0.000	0.000	0.000	0.000	-0.067	-0.034
Skewness	1.406	0.000	4.016	0.000	0.197	0.000
Kurtosis	3.915	-0.367	28.002	-0.288	-0.120	0.824
p(Skew Normality)	< 0.001	0.998	< 0.001	0.998	0.473	1.000
p(Kurt Normality)	< 0.001	0.968	0.001	0.974	0.826	0.133
p(Lilliefors Normality)	< 0.001	< 0.001	< 0.001	< 0.001	0.012	0.018
p(Paired 1-tailed t)		< 0.001		0.05	0.05	0.03
p(Wilcoxon)		<b>0.04</b>		<b>0.04</b>		<b>0.05</b>
p(Sign test)		0.18		0.08		0.26

**Table 2.** Summary statistics for distributions of difference scores for matched-pair data sets ((nc – tc) scores for each subject in each scenario).

**Statistical Issues**

Data frequency distribution normality emerged as a critical concern. A combination of phenomena led to this abnormality and managing it became a serious statistical consideration. Table 1 summarizes the raw-score findings.

The substantial number of no-error runs skewed the raw frequency distributions. *Rmin* was more normal, since it was not subject to any fixed initial limit, but *red alerts* and *niti* suffered predictably from the problems endemic to any metric with a lower fixed limit.

A second bias toward non-normality was also predictable. Pilots logically tend to avoid other aircraft. This is a “hard limit” fixed around the boundaries of the protected zone. On the other hand, they also want to alter course as little as possible in the interest of efficiency. So amount of deviation-from-path is a “soft limit” based on multiple individual criteria, including experience and confidence.

As Table 2 shows, distributions of difference scores “behave better” than raw scores in these cases. Nonetheless, in all three measures the problem of normality remained. Only *rmin* passed even a preliminary, lax normality screen based on a z-score probability defined by

$$p\left(z = \frac{\text{Skewness}}{\text{Std. Error of Skewness}}\right) \quad (3)$$

(Bliss, 1967) and no measure passed the more-stringent Lilliefors test (Hollander and Wolfe, 1999). In the case of *red alerts* this was not surprising, since raw data values were constrained to be integers.

Data were therefore transformed to minimize third-moment skewness using a power function  $(X+k)^\lambda$ , *k* being a constant. Rasmussen (1989) addresses transformations and their effects on Type I error, and lobbies for a modest p-value correction (e.g. choosing .04 to maintain a nominal .05 level).

However, even though transformation produced major improvements in skewness, in our judgment the Lilliefors results were insufficient to justify using ANOVA for our data evaluation.

Therefore, the values we emphasize here are the result of Wilcoxon’s rank sum test. Values for the sign test are also reported for the sake of comparison. The sign test is the least

powerful of the standard non-parametrics, but also has the fewest distributional assumptions (Hollander and Wolfe, 1999).

**Main Effects**

According to the Wilcoxon results with the Rasmussen correction, there were main effects for *red alerts* and *niti* (p = .04). *Rmin* (p = .05) missed significance by such a small margin that we choose to discuss it as well.

There was superior collision-avoidance performance with commercial airline pilots using the conflict-probe CDTI. Pilots displayed fewer discrete separation failures, had greater average points-of-closest-approach and showed lower overall severity on the failures that did occur.

However, any judgment concerning effect should be tempered by an estimate of effect size. One such measure is estimated z-units (Glass and Hopkins, 1984), given by delta-hat

$$\hat{\Delta} = \frac{\text{Mean Difference}}{\text{Pooled Std. Deviation}} \quad (4)$$

The effect sizes of *reds/niti/rmin* (.30/.34/-.24) are seen to be somewhat modest using this criterion. This implies that other considerations, such as conflict severity and individual differences in resolution ability, should be of more future concern than conflict detection alone.

**Treatment Order Effects**

Treatment order effects are a problem of within-subjects designs. We counterbalanced the treatment order to control for this, but believed it critical to test for asymmetrical transfer, fatigue, practice, and Ebbinghausian “savings-in-learning” effects, since each subject experienced each scenario twice.

All sixteen pilots reported they had no knowledge that scenarios had been repeated. However, some performance effects may occur unconsciously, so statistical evaluation was conducted.

Performance scores were first sorted by presentation order. Regression analysis showed no statistically significant overall performance relationship with time on any measure. Although regression lines uniformly sloped in a direction consistent

with fatigue effects, the slopes for *reds/niti/rmin* (+.01/+.38/-.01) were not significant (2-tailed  $t(8)$ ,  $p = .39/.23/.28$ ).

To assess asymmetrical transfer, scores were sorted by “tc-first” vs. “nc-first”. Subsequent analysis showed no significant effects for the order of display presentation (*reds/niti/rmin* 2-tailed  $t(14)$ ,  $p = .25/.34/.24$ ). From this we may infer that, if proactive and/or retroactive inhibition effects were at work, at least they appeared in somewhat equal balance across display order, and thus did not constitute a causal alternative to the main effects previously described.

### Catastrophic Separation Failure

*Rmin* and *niti* allowed a fine-grain examination of what could be called “catastrophic” separation failure. These results were not analyzed statistically, due to the fact that the definition of “catastrophic” can be considered arbitrary. However, sample values are presented in Table 3 to give a feel for the concept of how changing one’s threshold definition of catastrophe affected the number one would see in the data.

T	By Niti		T	By Rmin	
	nc	tc		nc	tc
1.0	27	15	1.00	26	16
2.0	25	14	0.80	20	10
4.0	22	9	0.60	12	5
8.0	16	7	0.40	5	1
16.0	7	2	0.20	2	1
32.0	2	1	0.10	2	1
103.0	1	0	0.06	1	0

**Table 3.** Incidence of experiment-wise separation failure as defined by a threshold value (T) for nc display and tc display.

Keep in mind that units in *niti* and *rmin* are normalized. So, for example, an *rmin* of .06 could correspond to either a pure lateral separation of about 556 m (1823 ft) or a pure vertical separation of 18.3 m (60 ft). So a “mathematical catastrophe” may or may not coincide with an experiential one. A lateral miss of 1823 ft would not feel as frightening as a vertical miss of 60 ft. But, according to the dimensions of the protected zone, they are equivalent.

We have no simple explanation for these catastrophies at this time. Some seemed due to pilot inattention. At other times, pilots appeared to mistrust the accuracy and reliability of the probe and spent time searching for conflicts no matter what the probe indicated. It is likely that a tendency to remain in conflict-detection mode diverted cognitive resources away from solution-finding.

### DISCUSSION

The most significant finding in this study is the overall error rate of 41% for commercial airline pilots in simulated free flight. This is error as the FAA currently defines error—discrete failure to maintain minimum mandated separation.

This elevated rate contrasts with the results of the Battiste, et al. (1997) study, and is more consistent with

Barhydt and Hansman (1997). Whether this was due to lack of a premaneuver solution-testing tool, or perhaps unreasonable scenario difficulty, or both was unclear. We therefore sought to assess whether or not our scenarios were simply unrealistically challenging. This was done by two means.

In our debriefing questionnaire we asked if pilots generally felt able to avoid the aircraft conflicts they had encountered. The mean score on a 7-point scale was 5.9 (s.d. 1.0). This supported the notion that the scenarios were challenging, but not impossibly so.

Additionally, we asked the pilots point-blank if, based on their experience, they thought the scenarios were unrealistic. Without exception, the answer was that they were very challenging, but not unlike situations they themselves had encountered.

We therefore conclude that free-flight safety will be very much a function of future technology. A conflict probe can help—but cannot be relied on as the sole aid to tactical collision avoidance. Conflict *detection*—while a challenging task—is not as challenging as conflict *resolution*. Tactical free flight under high-stress situations can be expected to be highly problematic unless effective conflict detection and resolution aids are part of the technology available to all. Further research is necessary to ensure development and testing of these aids. From a human factors standpoint, they will need to calculate and make salient the information most relevant to both conflict detection and resolution.

### ACKNOWLEDGEMENTS

This work was partially funded through FAA grant 93-G-048, Dr. Tom McCloy, Technical Monitor. We gratefully acknowledge the contributions of Dr. Kip Smith, Dept. of Psychology, Kansas State University, and of Mr. George Sweeney and Mr. Guy Smith, without whose help this work could not have been brought to completion.

### REFERENCES

- Barhydt, R., and Hansman, R.J. (1997). *Experimental studies of the effect of intent information on cockpit traffic displays* (Tech. Rep. ASL-97-3). Cambridge: MIT, Aeronautical Systems Laboratory.
- Battiste, V., Johnson, W., Delzell, S., Holland, S., Belcher, S., and Jordan, K. (1997). Development and demonstration of a prototype free flight cockpit display of traffic in formation. *Proceedings of the 1997 SAE/AIAA World Aviation Congress*.
- Bliss, C.I. (1967). *Statistics in Biology (Vol. 1)*. New York: McGraw-Hill.
- Glass, G.V., and Hopkins, K.D. (1984). *Statistical methods in Education and Psychology (2nd ed.)*. Englewood Cliffs, NJ: Prentice-Hall.
- Hollander, M., and Wolfe, D.A. (1999). *Nonparametric statistical methods (2nd ed.)*. New York, John Wiley.

- Knecht, W.R., and Hancock, P.A. (1997). Parameterizing a metric of midair collision risk. *Proceedings of the 41st Annual Meeting of the Human Factors and Ergonomics Society*, 9-12.
- Lozito, S., McGann, A., Mackintosh, M-A, and Cashion, P. (1998). Free flight and self-separation from the flight deck perspective.
- NASA Ames Research Center. <http://atm-seminar-97.eurocontrol.fr/lozito.htm>
- Poulton, E.C. (1982). Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. *Psychological Bulletin*, 91(3), 673-690.
- Rasmussen, J.L. (1989). Data transformation, Type I error rate and power. *British Journal of Mathematical and Statistical Psychology*, 42, 203-213.
- RTCA (January, 1995). *Report of the RTCA board of directors' select committee on free flight*. Washington, DC: RTCA, Incorporated.
- Scallen, S.F., Smith, K., and Hancock, P.A. (1996). Pilot actions during traffic situations in a free-flight airspace structure. *Proceedings of the 40th Annual Meeting of the Human Factors and Ergonomics Society*, 111-115.
- Smith, K.C.S. (1998). *Shared decision-making in the National Airspace System* (Grant No. 93-G-048). Washington, DC: Federal Aviation Administration.
- Smith, K.C.S., Lewin, J.E.K., and Hancock, P.A. (1996). The invariant that drives conflict detection. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics: Transportation Systems*. Aldershot U: Ashgate.