TRANSPORTATION HUMAN FACTORS, *2*(2), 99–126

Workload and Reliability of Predictor Displays in Aircraft Traffic Avoidance

Christopher D. Wickens, Keith Gempler, and M. Ephimia Morphew
University of Illinois

In 2 experiments we describe the relevance of aircraft predictor information to the availability and deployment of visual attention. In both, airplane pilots fly a simulator in which flight path prediction is given bearing on the future state of their own aircraft and of a second traffic "intruder" aircraft that they must maneuver to avoid. The cockpit traffic display on which this information is depicted is an integral component of the concept of *free flight* or pilot self-separation. In Experiment 1 we show that added layers of predictive information improve performance, reduce mental workload (as subjectively measured), and that added complexity of the visual display thus resulting does not increase the inferred measure of head downtime (secondary-task performance). In Experiment 2 we examine the consequences to performance and visual attention if prediction is occasionally in error. We adopt the hypothesis that trust is related to the relative allocation of attention between the predictor symbol and the raw data of actual aircraft state. Such unreliability damages performance to some extent, particularly when the unreliable predictor forecasts more complex conflict geometry. This cost reveals the substantial allocation of attention to the predictor symbol. However, pilots, knowing the level of unreliability, appear to be relatively well calibrated in their allocation of attention between the 2 information sources.

Predictive displays can provide considerable assistance to the human operator in transportation systems. This is particularly true for systems with high inertia or long lags in their response, such as the supertanker (VanBreda, 1999; Veldhuyzen & Stassen, 1977) and the wide-bodied transport aircraft (Gabriel, 1993) and is true with the displayed response of all aircraft as viewed on the air traffic controller's (ATC) display (Wickens, Mavor, & McGee, 1997). The source of this assistance emerges from analysis of the role of the human as a controller of lagged systems (VanBreda, 1999; Wickens, 1986). To make system position correspond with some required, time-varying command input (e.g., the center of the ship channel or middle of the flight path), the operator must execute a response now to match a command input that will not correspond to system position until some later time, a time that is equal to the system lag, which may be as much as a few minutes in the control of a large ship. Anticipation is not a function that people perform with high precision (Wickens & Hollands, 2000); it often requires estimating future input if this cannot be previewed, and it also requires running a mental model of the system dynamics, which can be a source of considerable workload (i.e., "If I provide this input to the control now, what will the output position and velocity be T seconds into the future?"). Furthermore, anticipation requires estimating what the desired target position will be at that

Requests for reprints should be sent to Christopher D. Wickens, Aviation Research Laboratory, University of Illinois, Willard Airport, ARL #1 Airport Road, Savoy, IL 61874. E-mail: cwickens@psych.uiuc.edu

future lag, which is sometimes visible in perspective (e.g., the channel, roadway, or runway ahead). Such future target information is sometimes called *preview* (Wickens, 1986; Wickens & Hollands, 2000).

Despite their apparent (and demonstrated) benefits, predictive displays also have at least two possible costs, which are the focus of the experiments we report on in this study. Experiment 1 addresses the workload associated with predictor displays, and Experiment 2 examines their reliability. Regarding workload, although predictive displays can offload the mental computation of prediction, thereby alleviating one source of resource demand, by definition they also provide added visual information to process. In many environments within the transportation industry, such processing may come at the cost of visual attention allocated elsewhere; for example, the driver who may need extra time to foveate and interpret a predictor of traffic patterns on the roadway ahead will spend that time with the eyes off the road, thereby creating a source of danger. Similar arguments could be made regarding the pilot in a crowded airspace. At the same time, a counterargument could be made based upon the task interference data described by multiple resource theory (Wickens, 1991; Wickens & Hollands, 2000) that the added perceptual activity imposed by processing the visual predictor demands the same resources as the cognitive activity of the mental computations of prediction, an activity that is relieved by the predictor. Furthermore, adapting the tenets of ecological interface displays (Vicente & Rasmussen, 1992), which assert that perception should ideally replace cognition in system design, one may argue that the resource demands of processing a predictor are actually less than those imposed by accurate mental prediction, and hence an overall benefit to workload (and reduction in task interference with other perceptual/cognitive tasks) should result. Thus one aspect of the research we report here provides an assessment of visual and cognitive workload associated with a predictor.

The second issue we examine concerns the accuracy or reliability of the predictor display. Prediction is, by definition, an inference about future behavior that for a variety of reasons may not be totally accurate. In fact for the prediction of vehicle, aircraft, or ship location, we can characterize three classes of factors that will influence the accuracy of prediction (Wickens, 1986):

1. The time into the future, known as the span of prediction. Accuracy is lessened with longer spans.
2. The inertia of the vehicle (or system) whose trajectory is predicted. Thus heavier vehicles with greater inertia, such as ships and wide-body aircraft, will have longer accurate prediction spans than lighter ones, such as automobiles and light aircraft.
3. The forces acting on (or expected to act on) the system. Thus spacecraft traveling in a vacuum will typically have longer accurate predictive spans than aircraft traveling through gusty turbulence; and both will have longer predictive spans if they are either uncontrolled or controlled by autopilots than if they are controlled by a human operator (who may be inclined to input an unanticipated control movement at any time).

Factors 2 and 3 interact in that systems with higher inertia will be more buffered from the effects of disturbances, particularly those disturbances of higher frequency.

Although the sources of unreliable prediction can thus be categorized, the human performance consequences of such unreliability are far less well understood because these depend jointly upon the extent to which the operator distributes attention between the current behavior of the vehicle (which we term the raw data) and the predictive symbol, and upon the extent to which the operator commits to some form of irreversible behavior on the basis of an incorrect predic-

tive symbol. This issue is a second focus of our research and does not appear to have been addressed by prior research in vehicle control.

The transportation system of particular interest in our investigation of predictor displays is the cockpit display of traffic information (CDTI), proposed for aviation (Ellis, McGreevy, & Hitchcock, 1987; Johnson, Battiste, & Bochow, 1999). The CDTI is an integral component of a proposed revolutionary change in the national airspace procedures, known as free flight, whereby pilots will be taking increasing responsibility for route selection and strategic maneuvers, responsibility that currently resides with ATC (Planzer & Jenny, 1995; Wickens, Mavor, Parasuraman, & McGee, 1998). (Such a system is not designed to replace ATC in most regions of airspace but to do so only in regions where pilots can engage in self-separation with little anticipated costs to safety and with clear gains in efficiency.)

A handful of studies have examined the value of predictor displays for CDTIs (Barhydt & Hansman, 1997; Hart & Loomis, 1980; Johnson et al., 1997; Kreifeldt, 1980). Sometimes these displays are referred to as *conflict probes* (Barhydt & Hansman, 1997) in that they will project or probe the future behavior of the pilot's own aircraft (henceforth called *ownship*) and the traffic aircraft to assess if that behavior predicts a conflict. Such studies have established the value of predictors in the CDTI, just as other investigations have established the value of ownship predictors on electronic map displays (Gabriel, 1993) and flight-path guidance displays (Jensen, 1981; Ververs & Wickens, 2000) that are not used for traffic conflict avoidance, but rather for routine guidance and navigation. Barhydt and Hansman (1997) and Johnson et al. (1999) demonstrated the strong pilot preference for predictive information in a CDTI. However these studies have not undertaken a specific examination of the visual workload associated with using traffic prediction, and hence its implications for head downtime and, ultimately, for flight safety.

A second limitation of prior studies of predictive airborne traffic displays is that they have not examined predictor reliability issues: What are the consequences when a predictor is incorrect in its projection? On the one hand, imperfect reliability of an automation-based predictor of ownship behavior can be fairly easily tolerated because a major source of error variance in the predicted flight path will be the pilot's own choice to maneuver in a way that is different than he or she had earlier intended (and that the predictor assumed). For example if a pilot is flying straight, the predictor algorithm will infer the straight path to continue, and hence project a straight line forward on the display. If the pilot suddenly decides to turn, this straight line will no longer be an accurate projection of the future behavior of ownship (i.e., it will be unreliable). But the pilot, aware of his or her own turning decision, will not depend upon the straight line projection in planning, but instead will wait until the predictor path has bent to conform with the curvature of the turn.

In contrast, unreliability of predictions of traffic behavior (i.e., other aircraft) could have more serious consequences. (In a traffic-avoidance task, the predictor of the other aircraft could be referred to as preview because it defines the future target position.) These consequences could be serious because the pilot of ownship may have no prior knowledge of the intended departure from the predicted path. Hence the pilot of ownship could have planned (and initiated) a maneuver based upon the predictor element of the traffic and then suddenly found that this maneuver is no longer appropriate. If the pilot were monitoring the actual behavior of the traffic (rather than its predicted path), that is, monitoring the raw data, then a revision in plans could be rapidly implemented. But if the pilot instead were focusing attention heavily (or exclusively) on the predictor element and this element did not reflect the changed behavior, then problems could result. Such departure of predicted behavior from true behavior could result if the logic driving

the predictive display was not based upon the momentary dynamics of the airplane itself (e.g., current rate of turn), but rather upon a preloaded flight plan or on parameters loaded within the flight management system autopilots, and the pilot of the traffic aircraft decided to depart from those flight plans. Prior research does not appear to have examined this issue directly. Research (VanBreda, 1999) offers some support, however, for the projected difficulties with unreliable prediction on large ship predictors. VanBreda found that the benefits of such predictors in predictable and routine environments were greatly reduced if the environment itself became unpredictable. However, VanBreda did not directly degrade the accuracy of the predictor itself.

Both experiments that we report here employed the CDTI format that Merwin and Wickens (1996; Merwin, O'Brien, & Wickens, 1997) developed, presenting a coplanar view (top-down map and forward looking vertical view) with a 45-sec predictor attached to ownship. In certain conditions (Experiment 1) and with certain reliability levels (Experiment 2), a corresponding predictor extended from the intruder traffic aircraft because pilots were required to avoid a loss of separation with the traffic while minimizing the deviation from a prespecified flight path. In both experiments pilots also monitored a secondary task, simulating the visual workload demands of out-of-cockpit viewing. This task required detecting faint intensifications at the top of the display, flashes that could not be seen unless visual scanning was diverted upward away from the CDTI.

EXPERIMENT 1

Experiment 1 compared three conditions incorporating progressively more predictive information. On the basis of previous research (Merwin & Wickens, 1996), we anticipated that the additional predictive information would improve maneuver performance. However, the implication for both visual workload (assessed by the secondary task) and cognitive load (assessed by subjective measures) were uncertain and characterized the primary focus of the study.

Method: Experiment 1

Participants. Participants were 15 licensed flight instructors (all male) from the University of Illinois Institute of Aviation and received $5 per hour for their participation. The mean number of flight hours for all participants was 341 hr with 80.3 instrument flight hours.

Simulation flight dynamics and apparatus. The part task flight simulation was run on a Silicon Graphics 4D/30 Super Turbo workstation and viewed on a Silicon Graphics 20-in. color display. The display screen resolution was 1280 × 1024 pixels and was run at a frequency of 60 Hz. The simulation allowed participants to control ownship's airspeed, altitude, and heading, controlled through a flight stick located on the right-hand side of the Silicon Graphics workstation. The flight stick allowed maximum pitch and bank angles of ±5° and ±30° respectively to preclude any extreme maneuvers aimed toward evading impending traffic conflicts. Speed control was maintained through the flight stick as well, with increased speed (at a constant rate) resulting from pushing the button on top of the flight stick and decreased flight speed corresponding with pressing the trigger. The maximum speed change capability was ±150 Kts, which translated to a maximum flight speed of 475 Kts and a minimum of 175 Kts. Although pilots were provided with the capability of using speed control as a means of managing traffic con-

flicts, they were instructed to deviate from the prescribed speed of 325 Kts as little as possible. The same was instructed for ownship's prescribed altitude (10,000 ft) and heading values (toward the waypoint). Light turbulence was programmed into the simulation, causing ownship at times to drift slowly from the prescribed heading and pitch angles if active control was not maintained. The flight simulation dynamics were cross-coupled as is characteristic of real flight, so that banking the aircraft caused a pitch down attitude and pitching downward would increase airspeed.

Task and simulation. Pilots flew a series of 6 missions throughout 2 separate experimental sessions (3 missions per session). Each mission comprised 10 consecutive (1–2 min) flight scenarios. In each scenario, pilots were required to fly to a designated navigational waypoint located directly ahead of ownship without coming into conflict with the intruder traffic aircraft located in their airspace while maintaining the prescribed flight parameters to the greatest extent possible (heading, speed, and altitude). In every mission, 9 of the 10 flight scenarios were designed as conflict scenarios requiring the pilot to execute avoidance maneuvers to resolve an impending traffic conflict (i.e., to avoid a loss of separation). The task required pilots to determine if the intruder's flight path would penetrate the protected zone around ownship, and if so, to use any means of maneuvering (including speed, heading, or altitude control) to avoid such penetration. Pilots were instructed to maneuver to minimize deviations in speed, heading, and altitude from their prescribed values. Ownship's designated protected zone was ±1,500 ft vertically and 3 mi horizontally.

Displays. A schematic of the general display format used by all participants is presented in Figure 1. The display format chosen for the experiment was a two-dimensional coplanar display, with a top-down and corresponding forward-looking view of the airspace surrounding ownship. The display included an ADI (attitude directional indicator) located in the top center region of the screen, which was used by pilots to maintain aircraft attitude (pitch and bank levels). The vertical strips located on adjacent sides of the CDTI represent the altimeter (right) and airspeed indicator (left) respectively. Visual scanning demands of the forward field of view (FFOV) or out-of-cockpit view were imposed by low contrast ellipses that appeared at random locations and times throughout each trial within the horizontal bar extending across the top of the screen. Contrast between ellipses and background was adjusted downward so that the onset of these ellipses could not be detected in peripheral vision, thereby requiring the pilot to scan upward for their detection.

The navigational display, used to represent traffic, portrayed a top-down (map) and forward-looking (profile) view of the pilot's surrounding airspace. The traffic symbology was overlaid on a grid of equispaced lines representing 5 nmi increments, with dots positioned at intervals of 1 nmi. The grid rotated with ownship to provide consistent spacing information of traffic symbology. The top-down view of the traffic display contained air traffic symbology consisting of ownship and intruder's aircraft icons, and a waypoint symbol representing the location of the pilot's destination. Dependent upon the display condition, the display contained predictor lines on both aircraft, and a threat vector stemming from ownship's predictor line. Figure 2 illustrates the three display conditions.

All display conditions presented ownship with a predictor line—a vector projected from the nose of the aircraft extending 45 sec into the future, which provided pilots with a graphic depiction of their aircraft's future position based on currently maintained parameters (heading, turn rate, vertical speed, and airspeed). In the baseline (BL) display condition, only ownship had a

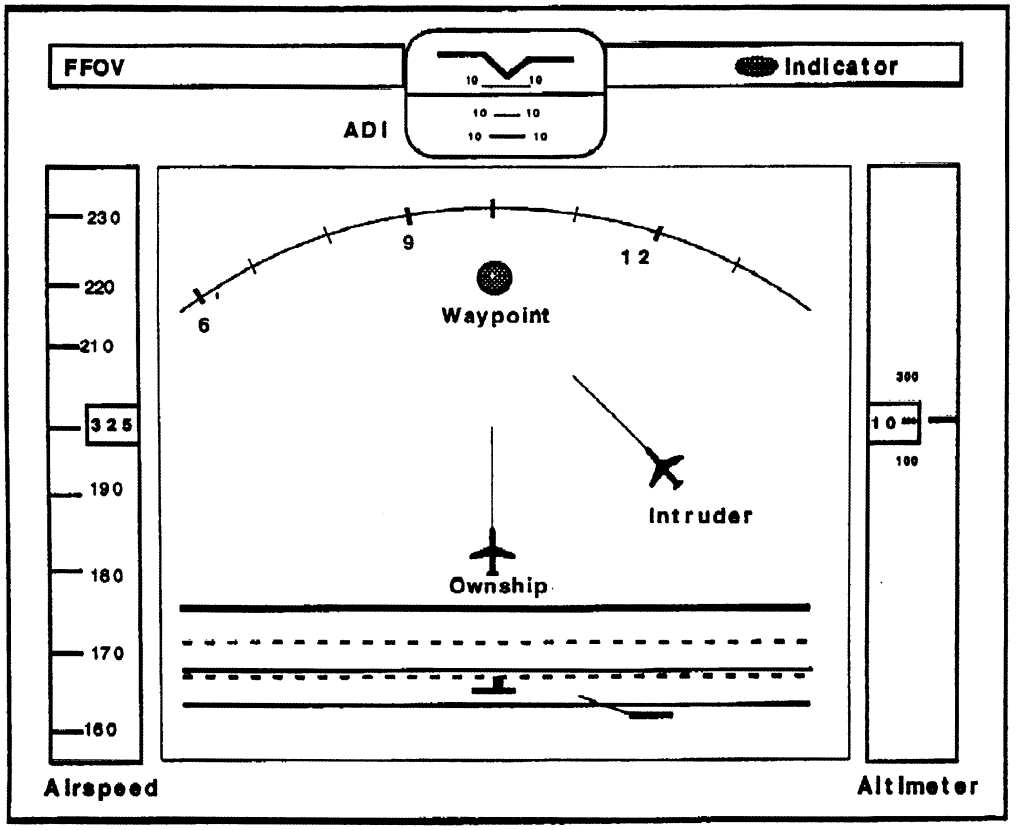
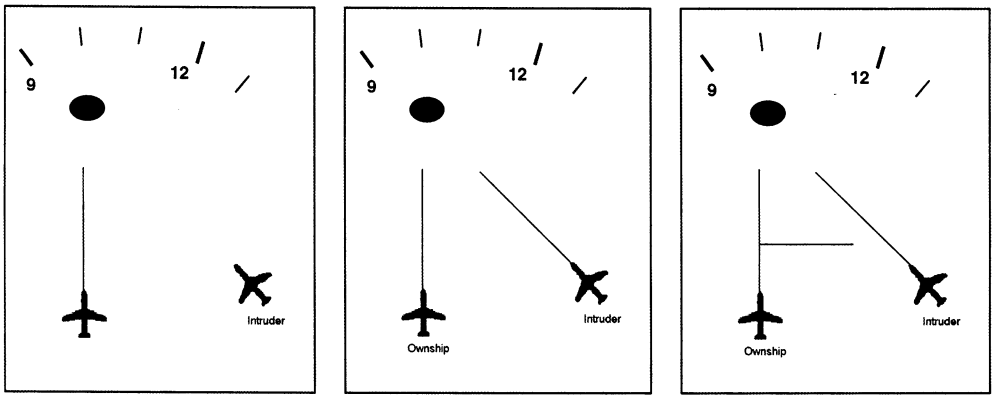


FIGURE 1 Cockpit display of traffic information used in the experiment.

predictor line, and no threat vector was displayed. The second condition (intruder predictor or IP) included predictor lines on both ownship and the intruder aircraft. The third condition (threat vector or TV) included predictor lines on both aircraft in addition to a TV emanating from a point along ownship's predictor line. This TV, which Merwin and Wickens (1996) developed as a planning aid, showed the relative bearing to the traffic aircraft that would exist at the point of closest passage, the projected separation at this point, and the predicted time until this point was reached. As a potential traffic conflict evolved, the endpoint of the TV would move closer to the IP line as the predicted minimum separation decreased. Additionally, the TV slides along ownship's predictor line, closer to ownship's aircraft symbol as the time to actual conflict or predicted point of closest passage decreases. Pilots were instructed to avoid contact between the TV's endpoint and the other aircraft's predictor line or aircraft symbol at all times because such contact would signal a predicted loss of separation (conflict) within 45 sec, or of an actual conflict if the TV connected ownship's symbol and the traffic. That is, when the TV contacted the other aircraft, it was within 1500 ft vertically or 3 mi laterally. The TV allowed pilots to perceive directly, rather than having to estimate, the current and future proximity of the intruder to ownship's protected zone. As the time to conflict between ownship and intruder decreased, movement of the TV along the predictor line toward ownship's aircraft symbol explicitly represented a countdown of the time to actual conflict (if a maneuver was not made). If the TV touched the IP line,



Baseline (BL) Condition | Intruder Predictor (IP) Condition | Threat Vector (TV) Condition

FIGURE 2 Three levels of prediction employed in Experiment 1.

a predicted conflict occurred and both aircraft would highlight indicating that the pilot was engaged in a predicted conflict.

Figure 3 depicts the distinction between the display symbology for a predicted conflict and an actual conflict. As Figure 3a demonstrates, when ownship's TV is not touching the IP line or aircraft symbol, no conflict is predicted. As Figure 3b illustrates, a conflict is predicted when ownship's TV is touching the IP or aircraft symbol. In this case, as indicated by the position of the TV along ownship's line, an actual conflict will occur in 20 sec if no evasive action is taken by ownship. Figure 3c depicts ownship in an actual conflict.

The forward-looking vertical view at the bottom of the coplanar display contained a set of parallel yellow horizontal lines representing ownship's current vertical protected zone boundaries (±1,500 ft above and below ownship). The dashed yellow horizontal lines represented the predicted vertical protected zone boundaries 45 sec into the future, computed on the basis of the current vertical speed.

Color coding was used for the symbology in the traffic display as a means of facilitating pilot perception of aircraft status states. The pilot's aircraft symbol and predictor line were magenta, whereas the intruder and its predictor line were gray. The TV was always orange. When ownship was in predicted conflict with the intruder, the two aircraft and associated predictor lines would highlight.

The secondary task was imposed by the FFOV indicator symbol superimposed on the horizontal strip extending across the top of the screen, which simulated the visual scanning demands of the FFOV or out-of-window view. The indicators appeared in randomly designated locations across the horizontal bar, at randomly generated times throughout each trial. Each indicator remained visible for a 15-sec period, or until noticed and acknowledged by the pilot by pressing the space bar on the flight simulator keyboard. Three or four FFOV indicator symbols were presented in each approximately 1.5-min trial. The pilot's task was to maintain his or her flight parameters (heading, airspeed, and altitude) while detecting and avoiding traffic conflicts with the CDTI, and maintaining attention in the FFOV region in the display, as would be expected in normal flight in good visual conditions. The appearances of the FFOV indicators were configured not to be detectable through the pilot's peripheral vision, but rather had to be directly foveated to be detected. Response time (RT) and accuracy of onset detection were recorded.

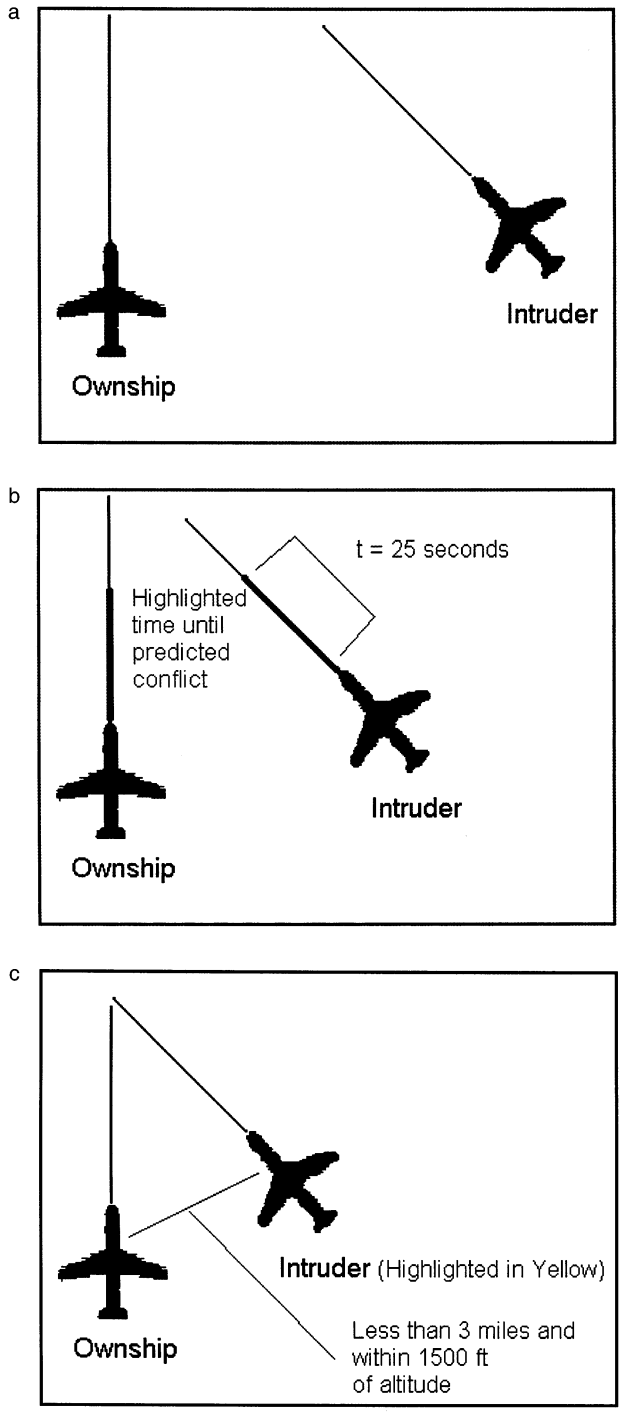


FIGURE 3 Distinction between the representation of (a) no conflict, (b) predicted conflict, and (c) actual conflict.

Experimental design. A 3 × 3 × 3 factorial, within-subjects design was used. The factors of interest were display type (BL, IP, TV), vertical traffic geometry (ascending, level, descending), and longitudinal geometry (45°, 90°, and 135°). The order in which pilots saw the three display types was counterbalanced across Sessions 1 and 2. The order of the different conflict geometries was randomized.

The intruder approach geometries with respect to ownship were varied to ensure exposure to a variety of traffic patterns, including three vertical geometries (ascending, level, and descending), and three longitudinal geometries (45°, 90°, and 135° = overtaking, crossing, and approaching, respectively). Each conflict could approach from either the left or right side of ownship. Each pilot saw every possible combination of intruder vertical and lateral geometry (including left and right approaches). Although an approximately equal number of intruder approaches were from the left and the right, this was not classified as an independent variable. Together, these geometries combined to produce 18 ways in which ownship was approached by the intruder aircraft.

Procedure. Pilots participated in two sessions on two days, each lasting approximately 1.5 hr. In Session 1, they were read instructions and shown illustrations relating to the task and display symbologies used in the experiment to familiarize themselves with the simulation. Pilots then flew 12 practice scenarios during which they familiarized themselves with the three displays, the display symbology, and the flight task. Upon successful completion of these practice trials, pilots began Session 1 and completed the first 3 missions, one with each display type.

In Session 2, pilots flew an additional set of six practice trials and then completed the last three missions, encountering the three display types in the reverse order to that encountered in Session 1. Upon completion of each mission in Session 2, participants were administered the NASA task load index (TLX) multidimensional subjective workload scale (Hart & Staveland, 1988) for each display type (BL, IP, TV). They then completed a postexperiment questionnaire that queried their preferred traffic avoidance maneuvers and strategies, as well as their preferred display type. They were instructed to balance their visual attention between outside scanning and instrument monitoring (including the CDTI) as they would in normal flight, and to avoid both actual and predicted conflicts while minimizing any deviations from the three prescribed flight parameters.

Results: Experiment 1

Statistical analysis. In the following analyses we adopt a two-level criterion for significance, highlighting "marginally" significant effects ($.10 > p > .05$) as well as significant ones ($p < .05$). We do this for two reasons (see Wickens, 1998, for further discussion). First, we believe that statistical inference should not be classified dichotomously, but rather viewed as a continuum of strength of evidence. A two-level criterion scale helps eliminate this dichotomization tendency. Second, the standard .05-level criteria is one designed only to prevent Type I statistical errors (falsely disproving the null hypothesis; Loftus, 1996). Particularly in human factors research, where operational design decisions with potential safety implications may be based on the output of statistical tests, we feel strongly that Type II errors are just as serious as Type I errors. That is, falsely rejecting a safety-improving innovation (a Type II error, concluding no difference when there is one), is just as serious as falsely accepting a difference (improvement) that does not exist. Loosening the criterion from the standard .05 level is a safeguard against com-

mitting Type II errors. Finally to prevent this strategy from unduly proliferating the number of significant effects, we only describe "marginal" effects ($.10 < p < .05$) to the extent that these are consistent with and supported by other aspects of the data.

Primary task: Traffic avoidance. Figure 4 presents the effects of display on the mean percentage time spent in actual conflicts with the traffic. A repeated measures analysis of variance (ANOVA) revealed that this critical parameter of safety yielded a marginally significant effect of display type, $F(2, 28) = 3.01$, $p = .09$, revealing a monotonic improvement in safety as more display information was provided (i.e., from BL to IP to TV). Figure 5 portrays a stronger effect of display type on the percent time in predicted conflicts, $F(2, 28) = 19.28$, $p < .01$. These are less serious conditions, but ones that, if they occur, would probably alert ATC in a free-flight scenario with the need to intervene. The data reveal that each added predictor display feature produced a reliable reduction in the number of predicted conflicts. No interactions between display type and the conflict geometry occurred for either of the two safety parameters.

The flight trajectories were analyzed by assessing the root mean squared (RMS) deviations of the maneuver away from the commanded trajectories (i.e., initial target parameters) on each of the three controlled axes. Analysis of these data revealed a marginally significant effect of display type on lateral deviations, $F(2, 28) = 3.10$, $p = .10$, with a pattern suggesting that the TV display induced more lateral maneuvering than did either the BL or IP displays. This pattern was reinforced by the finding that the TV display also yielded longer trajectories than the other displays (and significantly longer ones than the IP display; $F(1, 14) = 4.46$, $p = .05$, and induced marginally a greater amount of lateral control displacement than did the IP display, $F(1, 14) = 3.20$, $p = .10$). No evidence was found from the flight-path deviation measures that the display

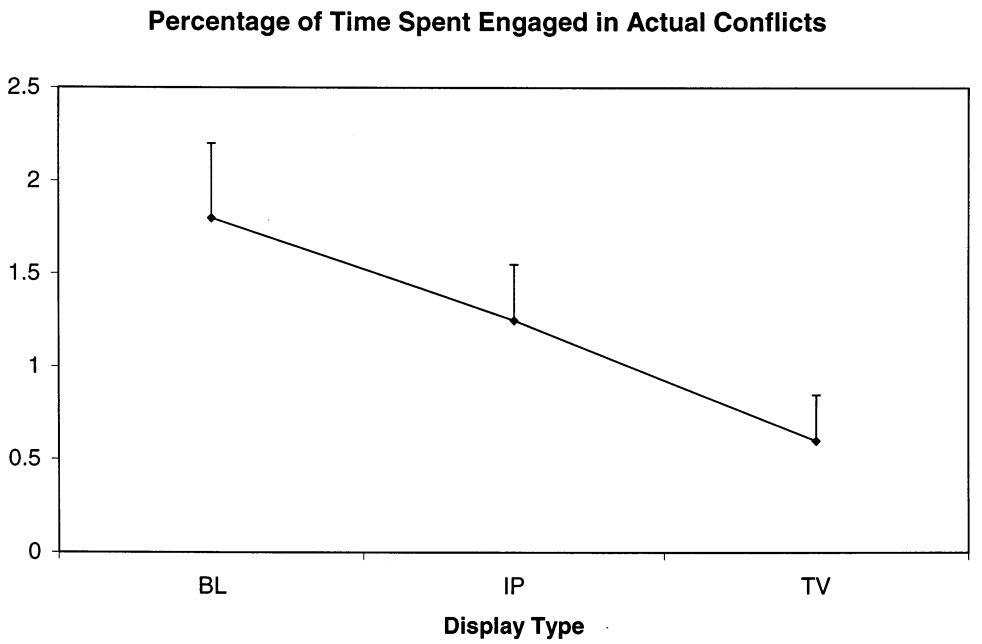


FIGURE 4 Effects of display format on actual conflict frequency. Bars depict one standard error.

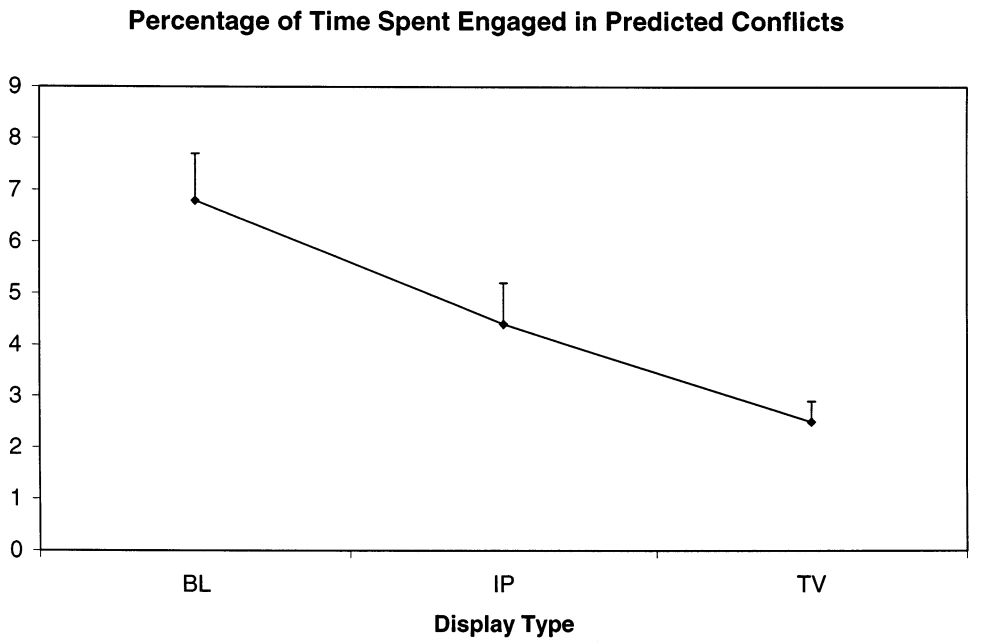


FIGURE 5 Effects of display format on predicted conflict frequency.

type altered the amount of vertical maneuvers selected. The data suggested that display type did not influence the amount of airspeed maneuvering.

Primary task: Subjective workload analysis. Analysis of the overall NASA TLX workload score (average across the 6 subscales) as a function of display type, revealed a pattern very similar to that shown by the safety measures, $F(2, 28) = 3.08$, $p = .08$. That is, a monotonic and marginally significant trend toward lower workload was found as progressively more display information was provided. Figure 6 breaks the workload effects down into the separate subscales and reveals two important characteristics. First, the major reduction in workload across all scales is evident between the BL and IP displays with few differences observed between the IP and TV displays. Second, the exception to the first characteristic is in the mental demand scale (Scale 1), which separates all three conditions by approximately equal amounts. Indeed a separate ANOVA conducted only on this mental demand subscale revealed a highly significant main effect of display, $F(2, 28) = 9.39$, $p < .01$. Significantly, in the mental demand scale, larger differences (greater variance accounted for) were observed between the BL and IP display, $F(1, 14) = 7.0$, than between the IP and the TV display, $F(1, 14) = 3.6$. This difference will have some importance in interpreting the results.

Secondary task: Event detection. Mean RT for the monitoring event-detection task was approximately 4.1 sec, indicating that pilots spent a good deal of their time head down, hence producing a substantial lag in detecting the FFOV secondary task events. (We assumed that RT to these events in single task conditions would be much less than 1 sec because the onsets were well in excess of threshold in foveal vision.) This time did not differ among the three display con-

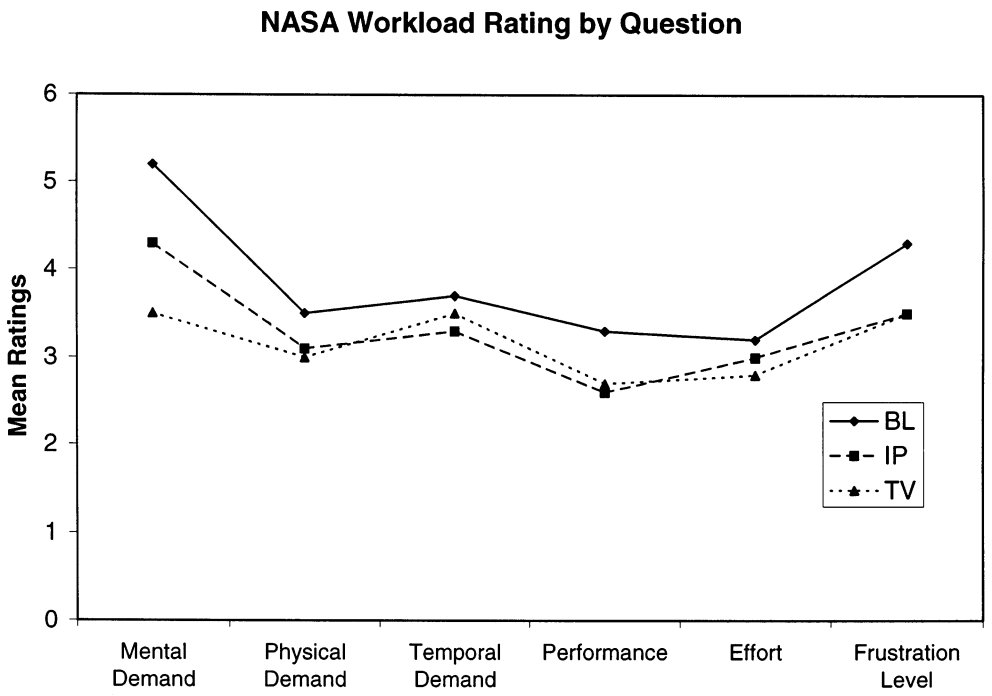


FIGURE 6 Effect of display format (the three lines) on the different subscales of the NASA task load index subjective workload rating.

ditions. Mean accuracy was also equivalent between conditions, with an average hit rate of about 0.70. Although false alarm rate showed a monotonic trend to increase with more display information (i.e., from BL to IP to TV), this effect did not approach statistical significance ($p > .10$).

Conflict geometry effects. Two relatively weak effects of conflict geometry were found on safety measures. Descending traffic produced fewer predicted conflicts than either level or ascending traffic, $F(2, 891) = 2.76$, $p < .06$, and overtaking (45°) conflicts produced fewer actual conflicts than crossing or approaching conflicts, $F(8, 891) = 2.38$, $p = .10$. Morphew and Wickens (1998) give further results of questionnaire and pilot comment results.

Discussion: Experiment 1

The primary objective of Experiment 1 was to examine the workload implications of levels of predictive information for traffic depiction. The results reveal that the original design intentions Merwin and Wickens (1996) used to create the display symbology according to cognitive engineering principles was successful. Such principles are based upon providing information in a format that directly serves the cognitive characteristics of the pilots' task and does so in a way that replaces cognitive operations of prediction, extrapolation, and spatial computation with perceptual ones (Vicente & Rasmussen, 1992). In this context, the impact of the two predictive display components can be examined, the addition of IP information and the addition of the TV.

Figure 7 summarizes the primary results of the study in terms of the implications of adding each "layer" of predictive information. As Figure 7 reveals both of these additions supported safer performance, marginally so in the case of actual conflicts, and more strongly in the case of predicted conflicts, which are frequently observed events, and hence events of greater statistical power. Further analysis however revealed that the nature of the support provided by each predictor element was slightly different. Providing the threat vector allowed or encouraged pilots to fly slightly different maneuvers using a greater amount of lateral deviations and lateral control. In contrast, providing the intruder predictor information had no influence on control or maneuver strategy behavior, but presumably allowed pilots to do the same job better than in the BL condition. This difference in the effects of the two display augmentations was reflected in differences in assessed workload. Figure 6 reveals the pronounced drop in nearly all workload aspects created by providing the IP. However, the figure also reveals that except for the mental demand scale, any further workload reduction that might have been provided by the TV was offset by either the added control activity that was induced to generate the greater lateral deviations or by the added perceptual effort required to process the (very useful) visual information offered by the TV.

Some evidence for the perceptual workload effect is also revealed by the fact that the added predictor "layers" produced no benefit for the secondary monitoring task (i.e., in terms of reduced RT). Furthermore, to the extent that false alarms may be viewed as a negative manifestation of accuracy, a hint of a systematic accuracy decrease was found on the secondary task as more predictive information was added to the primary task traffic avoidance display.

In Experiment 1, visual workload was only inferred by assessment of the secondary task response latency, rather than directly measured by visual scanning behavior. In this regard it is important to note that the inferred head downtime (4.1 sec) observed in the current experiment corresponds quite closely to the "mean first passage time" (Moray, 1986) away from the forward view (4.5 sec) measured directly from visual scanning behavior in a high-fidelity flight simula-

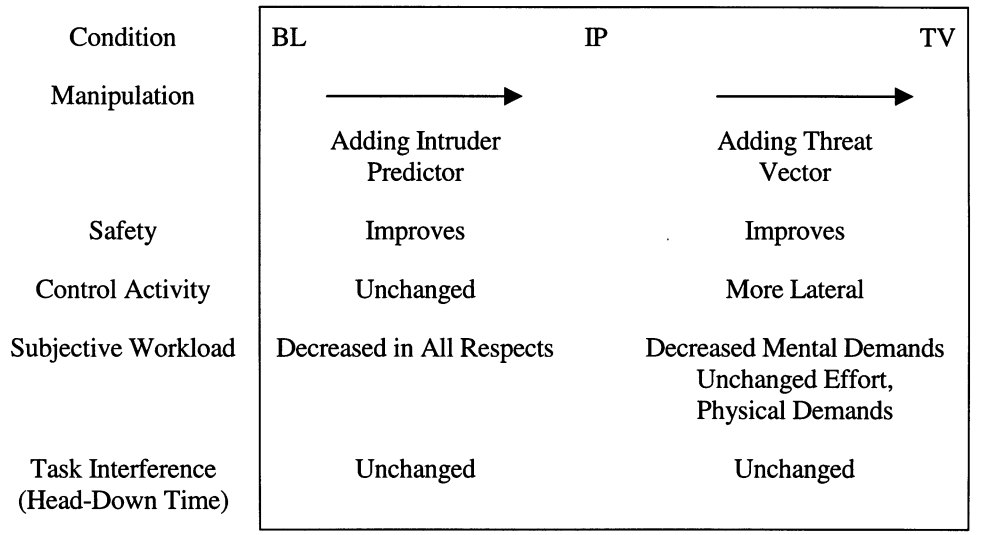


FIGURE 7 Effects of adding layers of predictive information in Experiment 1.

tion, employing a similar CDTI and experimental procedures to that used here (Wickens, Helleberg, & Xu, 2000). Such convergence of values provides some degree of confidence that the secondary task used here mimicked the visual attention demands of the pilot in out-the-window scanning in a higher fidelity flight task.

Overall then the results imply that adding layers of reliable predictive information supports conflict avoidance maneuvers and reduces mental or cognitive workload as revealed by subjective ratings. But a possible shift from lower cognitive load to increased visual processing with the added display information of the TV may mitigate any possible benefits for increased visual attention allocated to outside monitoring.

EXPERIMENT 2

Experiment 1 presented entirely reliable predictor information. As noted earlier, however, many scenarios are possible in which the prediction of traffic behavior is not fulfilled by subsequent events. Here we distinguish between imprecision limits of the predictive span, in which the eventual aircraft path does not entirely lie along the predicted path, particularly at longer intervals, and what we refer to as catastrophic failures in which a fundamental inflight shift of the path is undertaken, but the shift is not available to the device generating the predictive input. Figure 8 shows two examples. In the top part of the figure, the aircraft predictor is indicating a descent while the aircraft remains at a level altitude. In the bottom part, the aircraft predictor indicates a direct path, but the aircraft actually turns left while the predictor remains at a constant heading. Such dissociations are possible when different sources of information are used to derive current position (e.g., radar, global positioning system), and future intent (e.g., flight management system computers). In other contexts, we have referred to this behavior as a catastrophic failure (Merlo, Wickens, & Yeh, 1999). In this context we examine pilot workload and performance when such failures in inferring intent occur relatively infrequently (1 trial in 6) producing a predictor reliability of 0.83. Our interest is in the consequences to behavior when such unreliability is imposed. On the infrequent error trials, does (unwarranted) trust in the predictor lead pilots to ignore the raw data of the actual traffic intruder flight path and fly into a difficult situation? Alternatively do pilots, upon learning of the unreliability, retain a healthy degree of calibrated trust allocating relatively more attention to the raw data and less to the predictor than they would if the predictor were totally reliable, but hence offsetting some of the predictor benefits on the more frequent correct trials.

We measured trust explicitly (e.g., Kantowitz, Hanowski, & Kantowitz, 1997; Lee & Moray, 1992) by asking pilots to estimate the reliability of the predictor. Following the lead of previous researchers (Kantowitz et al., 1997; Merlo et al., 1999; Parasuraman, Mouloua, & Molloy, 1996), we also measured trust implicitly based upon the inferred usage or processing of the predictor versus current state information. We consider these two sources of displayed information to be automation and raw data, respectively. We also employ the secondary task indicator as an implicit measure of attention adjustment. To the extent that pilots become more vigilant of the raw data on the trial immediately following a predictor failure, we would expect to see diminished performance on the secondary monitoring task. Finally, interest in Experiment 2 is directed to one possible remediation of the consequences of overtrust or overreliance on partially reliable automation. Here we use a display technique, described following, to provide pilots with a graphic picture of the range of unreliability of the predictor symbol. In other research, graphic displays of spatial unreliability, not unlike the standard error bar in a graph, has facilitated performance

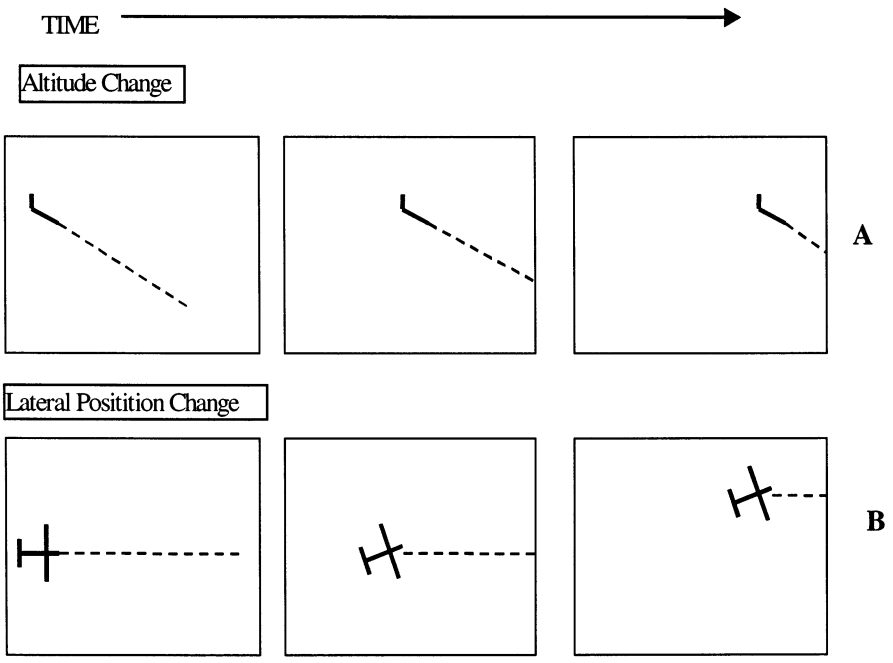


FIGURE 8 Two examples of three time frames depicting the behavior of an unreliable predictor. Top: unreliability in vertical behavior. The predictor indicates descent while the aircraft remains level. Bottom: unreliability in lateral behavior. The predictor indicates a left turn while the aircraft flies straight.

(Andre & Cutler, 1998; Kirshenbaum & Arruda, 1994; Merlo et al., 1999), although such efforts apparently have not been undertaken with predictive displays.

Method: Experiment 2

Participants. The 20 participants, 5 women and 15 men, were all licensed pilots with instrument ratings, most from the University of Illinois Institute of Aviation. They received $6 per hour for their participation. Seventeen of the 20 were instrument current and half were certified flight instructors.

The flight simulation dynamics and procedures were identical to those flown in Experiment 1, with the exceptions described following, and the display was identical to the TV display used in Experiment 1 (i.e., the full complement of predictor information on ownship and the traffic aircraft). Fifty conflict (C) trials and 10 nonconflict (NC) trials were conducted. To simulate a less than perfectly reliable predictor, 10 of the 60 trials had the intruder aircraft change heading or rate of climb or descent in a manner characterized in Figure 8. Heading changes were between 16.5° and 33.3°, and changes in pitch were from a climb or descent to level flight. Changes in trajectory occurred at unpredictable times between 14 and 38 sec (M = 24.3 sec) after the trial started. Changes from the predicted trajectory were not reflected by the automation generating the prediction. That is, as Figure 8 shows, the predictor line would continue to point in the direction of the original flight path. Of the 10 change trials, 6 were initially C trials and 4 were initially NC trials. The changes would either make the conflict easier to avoid (L = less

maneuvering required given the revised trajectory), or more difficult (M = more maneuvering required given the revised trajectory). The greatest focus of the study concentrated on the 8 trials in which more maneuvering was unexpectedly required.

The single line predictor display was identical to that employed in Experiment 1 (see Figure 1). The wedge predictor (W) display, depicting unreliability, included the same predictor lines on both aircraft but added two curved lines to the IP, one on either side of the predictor line in the shape of a wedge, that indicated an interval of possible future locations along the predicted path (W). The lines represented a 95% confidence interval of the aircraft's future location and the width of that interval increased with the forward distance from the traffic aircraft, indicating to the pilot that roughly 95% of the time the pilot could expect the intruder to be between these boundaries in the future. The change trial parameters were set so that this estimate would be approximately correct. The lines were generated using a parabolic function to give them their shape.

Experimental design. A 2 × 3 × 3 × 2 factorial mixed design was used. Display type (straight line or W) was varied between subjects, and vertical traffic geometry (ascending, level, descending) longitudinal geometry (45°, 90°, 135°) and trial predictor accuracy (correct, error), were varied within subjects. The 10 trials with an unpredicted trajectory change by the intruder, were quasi-randomly distributed within the set of 60 trials, thus rendering the predictor inaccurate for that trial and achieving an overall reliability of 83.3% for the entire series of 60 trials. The trajectory change trials were distributed in such a manner to appear random to the pilot. Pilots were assigned to each display group in a manner to ensure that both groups had a roughly equal distribution of flight experience. Both groups of subjects saw the exact same sequence of trials including 50 predictor valid trials and 10 predictor invalid trials.

The intruder aircraft approach geometries were varied to ensure exposure to a variety of traffic patterns, including three vertical (ascending, descending, level) and three horizontal (approaching, crossing, overtaking) geometries with approaches from both right and left of ownship.

Procedure. Pilots participated in one session lasting from 90 min to 2 hr consisting of four sets of 15 trials each. Before the session, pilots were read instructions and shown illustrations of the task and corresponding displays. They were specifically told that the predictor display was not 100% accurate. After instructions, the participants flew 10 practice trials with a reliable predictor to become comfortable with the displays. The experimenter sat in the simulation room with them during the practice session to answer any questions. Upon completion of the practice, pilots were invited to ask questions and then began the first set of trials. Between each set of trials, pilots were required to take a short (3–5 min) break before continuing with the simulation to avoid any fatigue effects.

Upon completion of the final set of trials, pilots were given a questionnaire that asked for their subjective estimate of the reliability of the predictor as an explicit measure of trust and their preferred avoidance maneuvers and strategies. Finally, pilots were asked for any additional comments, thanked for their participation, and paid for their time.

Results: Experiment 2

Prior to any analysis, data were examined for gross deviations from normality and outliers. No data points appeared to be outliers, and any transformations used are annotated in the analysis. All statistical tests were performed using Statistical Products & Service Solution, student version 6.1 for Windows.

Analysis of flight performance. Figure 9 depicts time in predicted conflict per trial with data for the line predictor shown in the top panel, and the W data shown in the bottom. The analysis revealed a main effect of approach geometry, $F(2, 663) = 11.00$, $p < .001$, and an interaction between predictor validity and approach geometry, $F(2, 663) = 20.34$, $p < .001$. The main effect of geometry reflects the large effect found in the interaction with predictor validity as seen in the right-most points of both panels of Figure 9. Here, the cost of inaccurate prediction trials was least on level trials, and was greatest on trials with a descending approach geometry; correspondingly, the cost of descending geometry was only revealed on the inaccurate trials. The predictor format, the line predictor in the left panel, and the W in the right had no effect.

Figure 10 depicts the measure of pilots' time in actual conflict per trial, which showed a main effect of validity, $F(1, 1188) = 8.09$, $p = .005$, of display format, $F(1, 1188) = 9.94$, $p = .002$, and of geometry, $F(2, 1188) = 23.80$, $p < .001$. The W display lead to more time in conflict than the line display ($M = 4.1$ and 2.2 sec, respectively). As with the time in predicted conflict, the effects of geometry and validity can again be best described in terms of their interaction, $F(2, 1188) = 31.13$, $p < .01$. Here, the cost of the invalid predictor (more time in conflict) appeared only on descending trials.

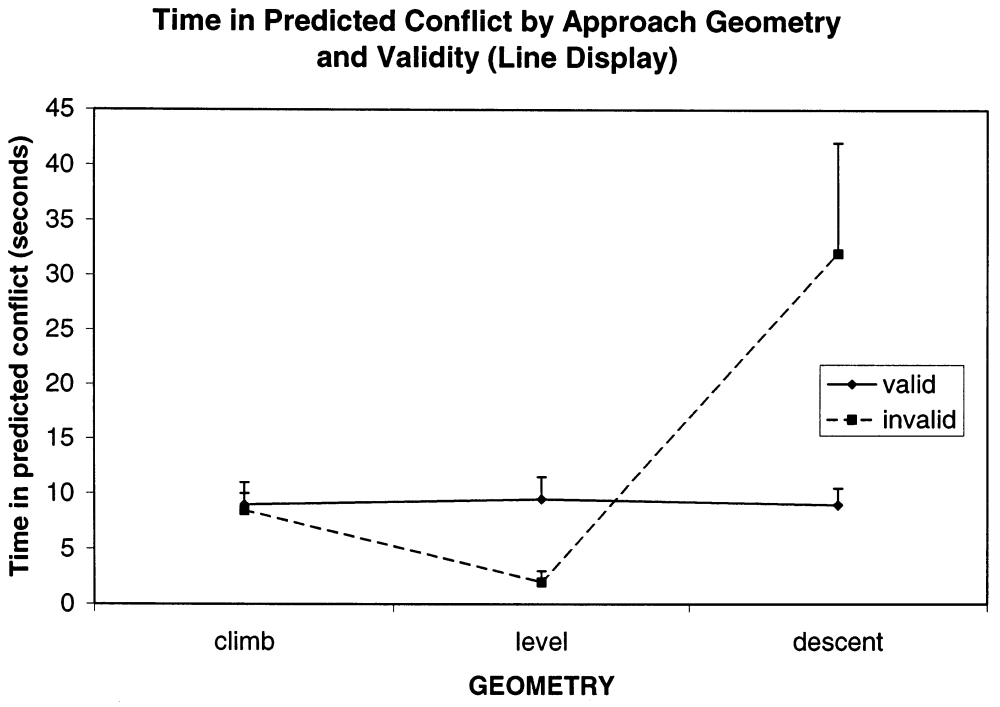
Figure 11 depicts the RMS deviation from the prescribed altitude of 10,000 ft in a scale of miles, reflecting vertical flight path efficiency. RMS deviation from altitude was analyzed using a log transformation to normalize the distribution. Analyses revealed main effects of predictor validity, $F(1, 1188) = 6.252$, $p = .013$, and of approach geometry, $F(2, 1188) = 12.49$, $p < .001$, as well as an interaction between predictor validity and geometry, $F(2, 1188) = 5.81$, $p = .003$. Climbing and descending traffic trials had a larger deviation than level trials (mean = 422 ft per trial). Once again, the main effect of validity can best be interpreted in terms of the interaction between validity and geometry, which reveals that invalid trials produced greater deviations from altitude than valid trials only when the intruder was descending. This pattern is similar to that observed with the safety measures of time in predicted and actual conflict.

RMS deviation from airspeed was analyzed using a logarithmic transformation. As Figure 12 shows, there was a main effect of display type, $F(1, 1188) = 22.26$, $p < .001$, with the W display having less deviation ($M = 29.3$ kts per trial) than the line display ($M = 45.3$ kts per trial). There was also a main effect of approach geometry, $F(2, 1188) = 3.70$, $p = .025$, which can best be explained in the context of the geometry by validity interaction, $F(2, 1188) = 9.18$, $p < .001$. This interaction is of the same general form as seen in the previous figures: the cost to invalid predictors was only seen with descending traffic geometry.

Figure 13 depicts the RMS heading deviation per trial from a prescribed heading of 360°. Once again, this RMS deviation measure was analyzed using a logarithmic transformation. Analyses revealed that there were main effects, of predictor validity, $F(1, 1188) = 24.76$, $p < .001$, and of display type, $F(1, 1188) = 4.46$, $p = .035$. The W display had a higher mean deviation than did the line display and invalid trials lead to consistently higher deviations compared to the valid predictor trials. No significant interactions were found in the data.

The time until the first maneuver was initiated was employed as a convergent measure of conflict problem difficulty. These data revealed that level traffic yielded shorter latencies than either climbing or descending traffic, $F(2, 1188) = 8.45$, $p < .01$. Also, showing the same pattern of interaction as many of the other variables, the descending geometry with the invalid predictor appeared to be particularly troublesome, yielding longest latencies, Interaction: $F(2, 1118) = 4.37$, $p = .01$.

The RT to the FFOV indicators can be viewed as a measure of the attention allocated to the primary traffic display. The analysis of these data, following a log transformation, revealed that



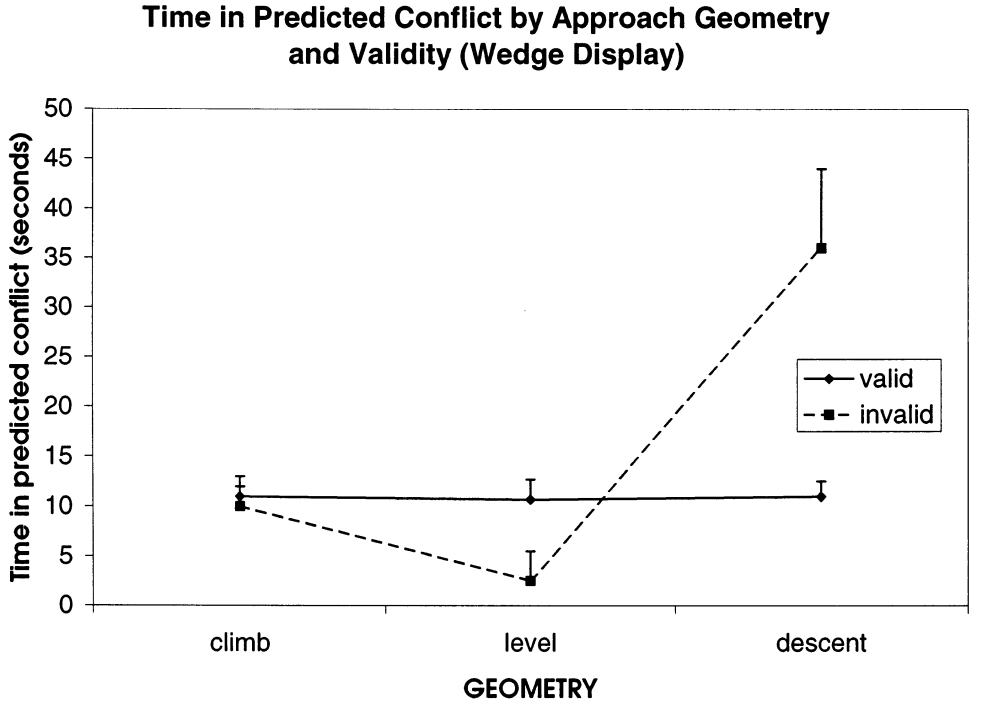
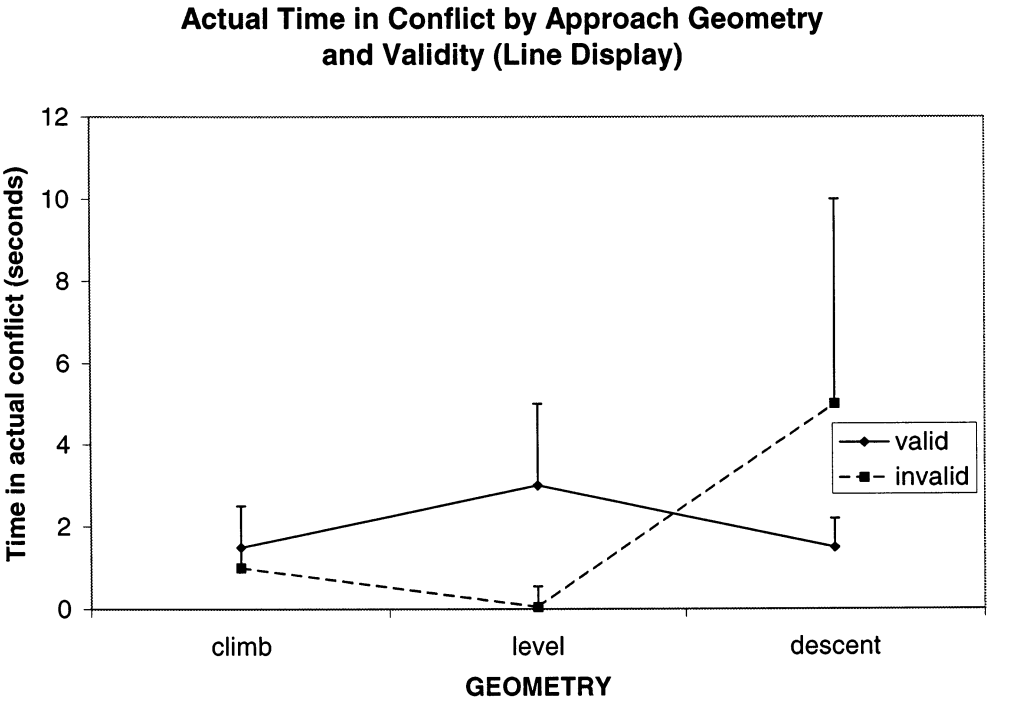


FIGURE 9 Time in predicted conflict measures (bar represents 95% confidence interval). Top panel is line display and bottom panel is wedge display.



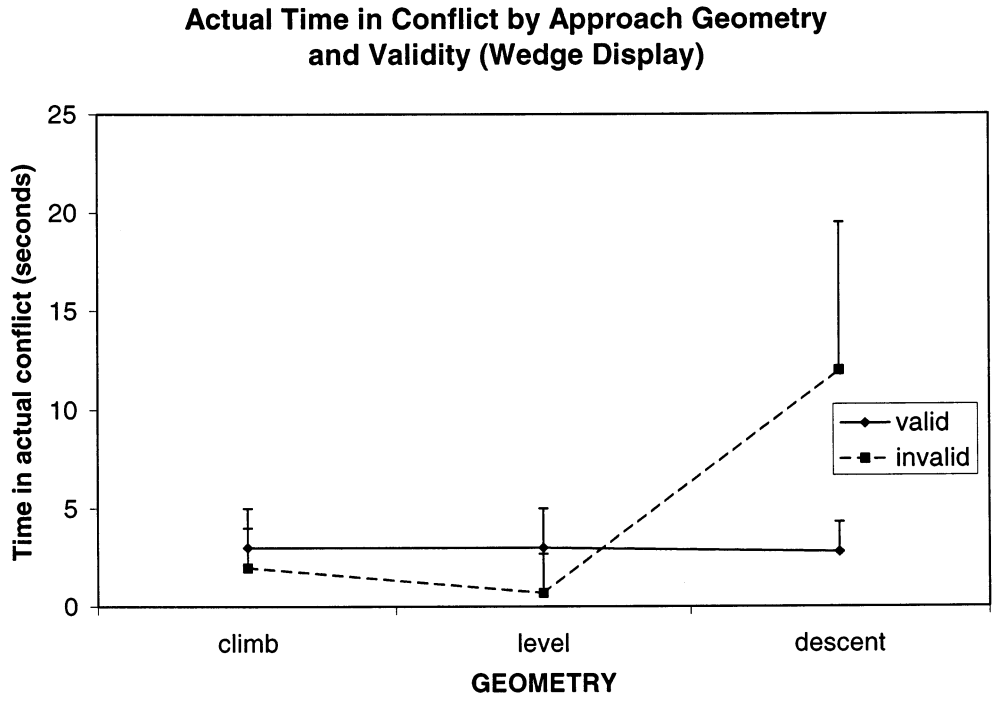
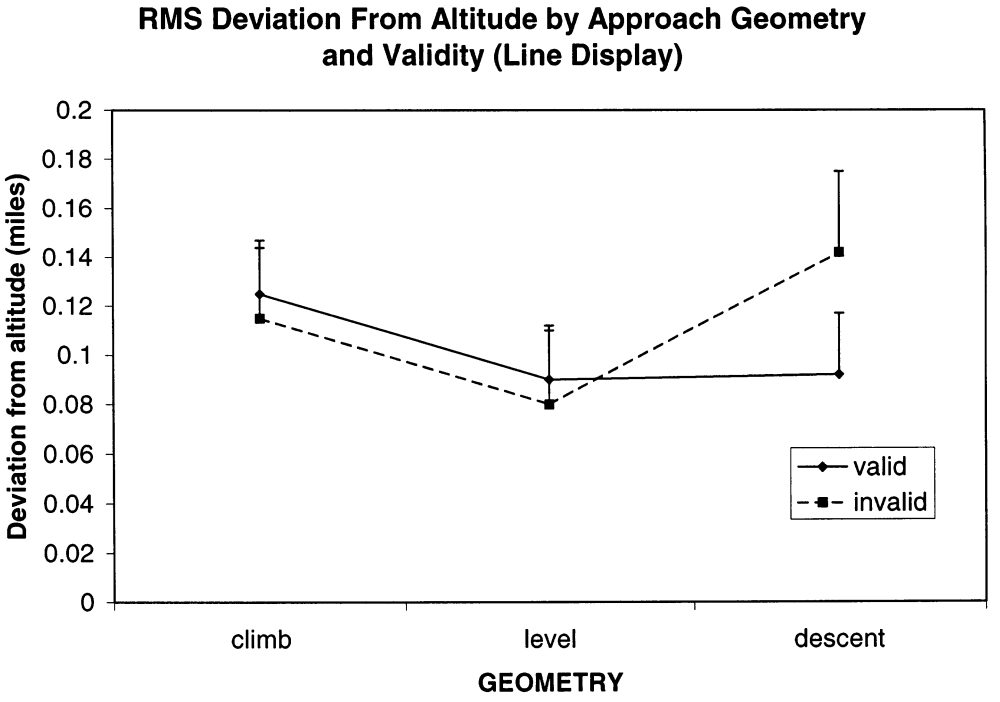


FIGURE 10 Time in actual conflict measures (bar represents 95% confidence interval).



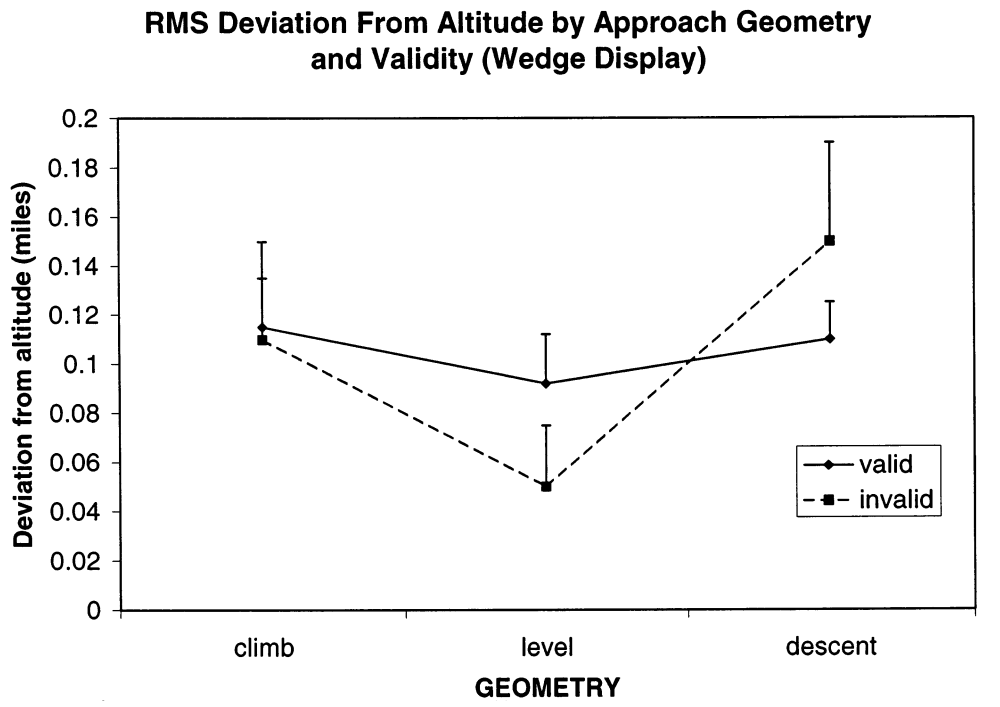
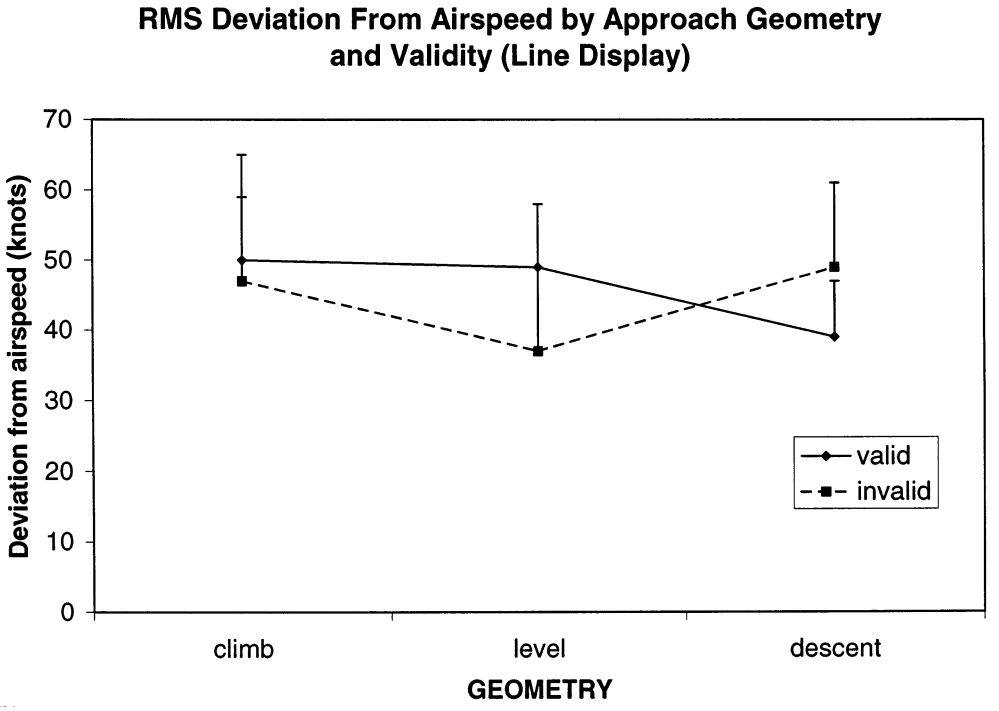


FIGURE 11 Root mean squared deviation of altitude measure (bar represents 95% confidence interval).



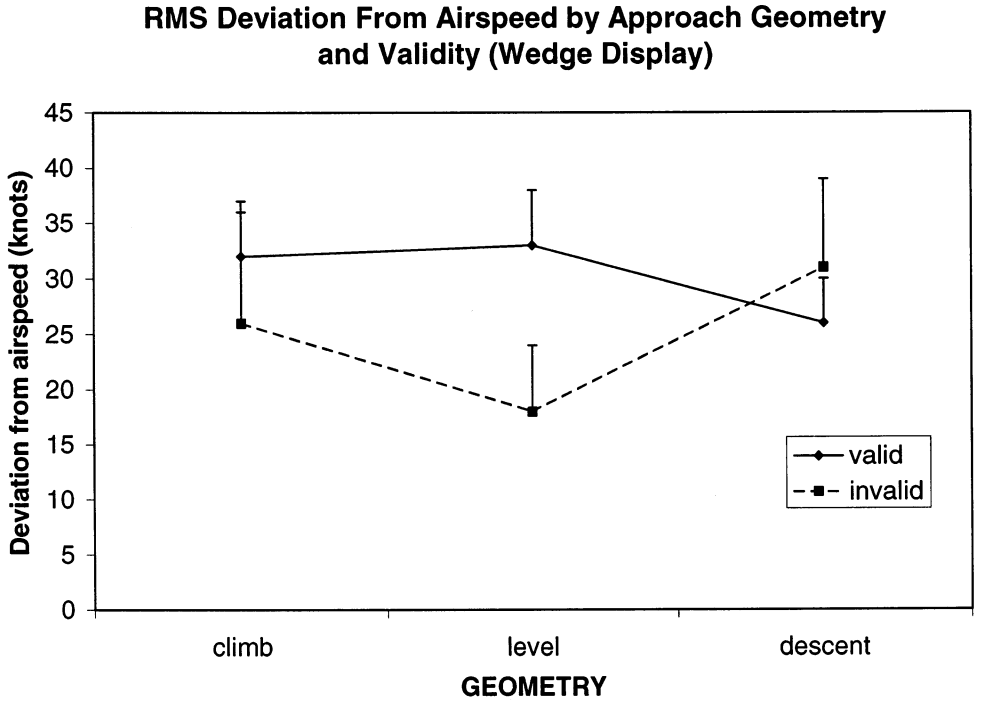
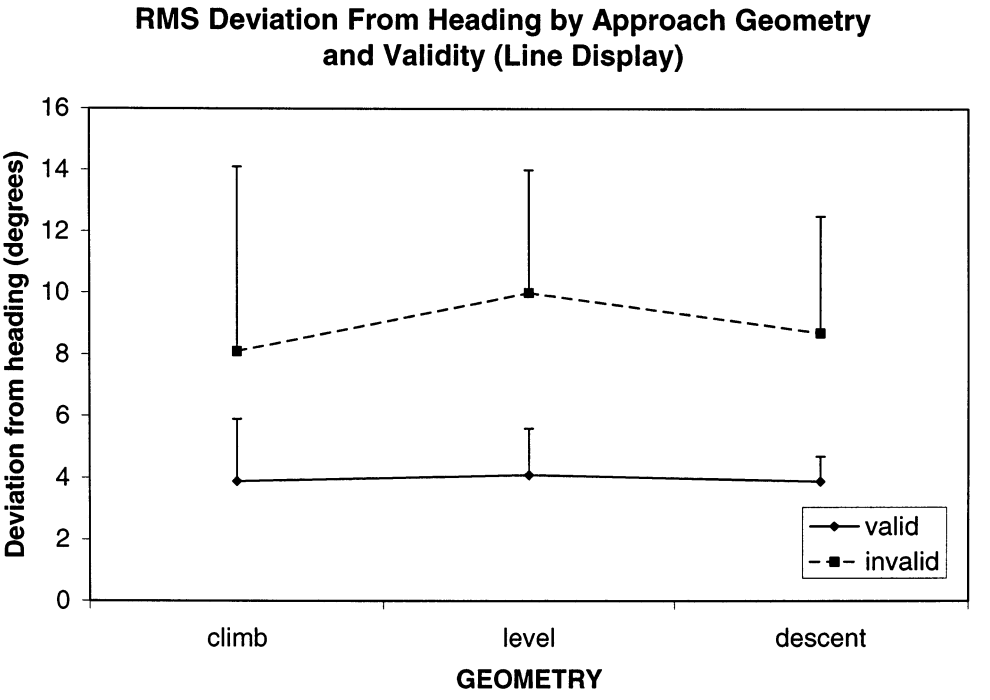


FIGURE 12 Root mean squared deviation of airspeed measure (bar represents 95% confidence interval).



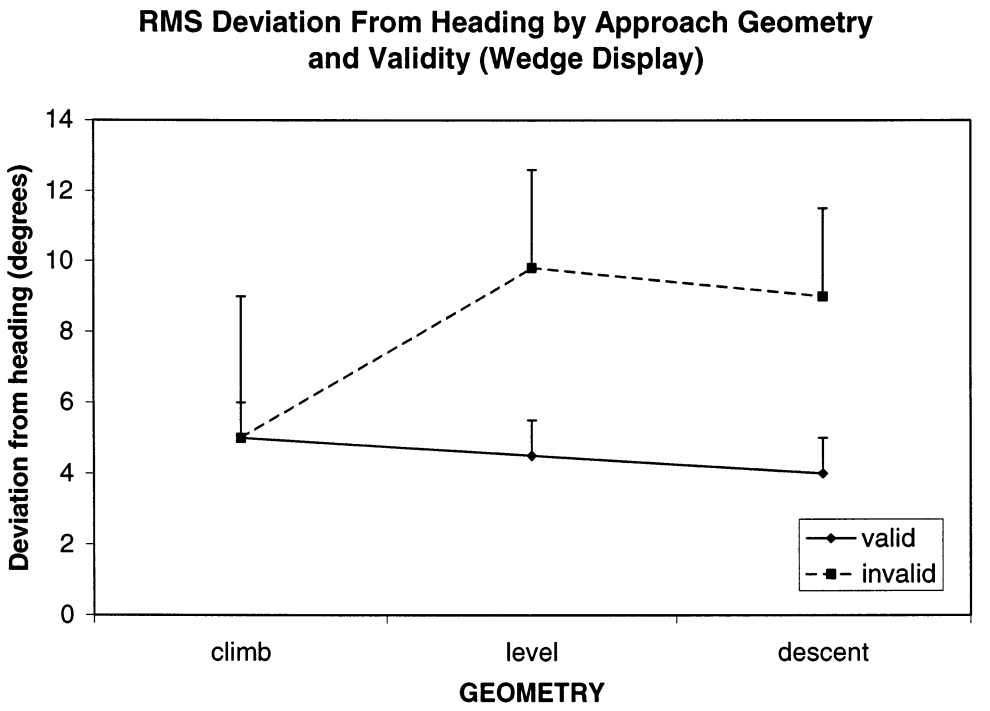


FIGURE 13 Root mean squared deviation from heading measure (bar represents 95% confidence interval).

trials on which the predictor was invalid lead to slower RTs (M = 4.81 sec) than the valid predictor trials (M = 4.52 sec $F(1, 1188) = 4.68$, $p = .03$). An effect of display type, $F(1, 1158) = 6.96$, $p = .008$, was found in which faster RTs were observed with the W display (M = 4.37 sec), than with the line display (M = 4.80 sec).

The analysis of FFOV task accuracy revealed a main effect of approach geometry, $F(2, 1188) = 3.69$), $p = .025$, indicating the greatest accuracy for level trials. The false alarm rate was very low in this measure (fewer than 5 false alarms throughout the 1,200 nonpractice trials) and was not a significant factor in the accuracy analysis.

The analysis of implicit trust was carried out by examining differences of safety measures, FFOV RTs, and time to first maneuver between the trial immediately prior to an invalid predictor trial and that immediately following an invalid predictor trial. In this analysis we only employed the invalid trials that made the maneuvering more difficult by decreasing time to loss of separation. We inferred that this trial type would induce the greatest—or most noticeable—loss of trust.

Although a faint trend for safer flying (less time in conflict) occurred following rather than prior to the error trial, this trend did not approach statistical significance. Correspondingly, neither measures of attention allocation, FFOV RT, or accuracy, showed a significant change as a consequence of the predictor error.

Finally, after the experiment was completed, pilots were asked to provide a number of subjective ratings. Many of these are discussed in Gempler and Wickens (1998); the most important for the current analysis was the pilots' estimate of the probability of predictor failure, that is, a rating of explicit trust. Pilots estimated an average of 7.9 failures, compared with the 10 that actually occurred, thus revealing some, but not perfect, calibration by slightly overestimating the reliability of the predictor automation.

Discussion: Experiment 2

Experiment 2 examined the effects of occasional inaccuracies (unreliability) in flight path predictive aids for a CDTI. We use the term *unreliability* with some caution because it normally has a negative connotation. We emphasize however that in many circumstances in a stochastic world, prediction simply cannot be perfect, and so occasional errors in a predictive device can and should be more readily tolerated than, for example, occasional errors in autopilots or in other automation relying upon deterministic environmental information. The errors of prediction Experiment 2 explores are simply the sorts that might occur if pilots depart from the preestablished flight paths that are themselves the source of the traffic predictor information. Our interest was in how such departures might influence the inferred allocation of attention between, and dependence upon, elements on the display. The results revealed that the effects of predictor unreliability can be best interpreted within the context of differences in the difficulty or complexity of the traffic pattern, so we discuss these issues first.

The evidence Experiment 2 provides, coupled with that from previous and subsequent research (Merwin et al., 1997; Wickens, Helleberg, & Xu, 1999, 2000), as well as an intuitive analysis of task complexity, suggests that traffic conflicts in which the intruder approached at the same altitude (level) imposed less workload than those in which the intruder was climbing or descending. In the current data, this was revealed by more accurate performance of the FFOV secondary task, faster initiation of the avoidance maneuver, as well as smaller deviations from the prescribed altitude during level trials. In three experiments (independent replications),

Merwin et al. (1997) and Wickens, Helleberg, and Xu (1999) had all observed greater time spent in predicted conflict with the nonlevel traffic trials; and these findings collectively are consistent with the task analysis of such trials, revealing that they require pilots to predict changes along three axes of flight (relative to ownship's movement), whereas level trials require prediction along only two. Similar evidence for workload differences in air traffic control between level and nonlevel flight was provided by Lamoureux (1999). (Interestingly, for reasons that cannot be fully explained, Experiment 1 provided no evidence for easier performance on level trials.)

If we therefore assume that the climbing and descending trials were more complex, it is equally plausible to assume that on these trials pilots would have placed greater reliance upon the predictive aiding to assist them in conflict avoidance because Experiment 1 showed such aiding to be useful and workload reducing. Correspondingly, greater dependence upon (and inferred attention allocation to) the predictive aiding would be anticipated to lead to more serious problems on the 1 trial in 6 when that aiding was in error. Indeed these results were observed, as Figures 9 and 10 show. In particular, Figure 9, portraying predicted conflicts, shows the V-shaped pattern on unreliable trials that mimics the workload costs shown by the FFOV accuracy data and by the maneuver task difficulty data as reflected in the time-to-initiate measure.

For reasons that remain less clear, the costs of unreliability appear to be more severe when the traffic was descending, than when the traffic was climbing. This difference is possibly related to the inherent tendency of pilots using this CDTI format to maneuver vertically in the direction of the traffic vector. That is, when traffic is climbing, the pilot descends; when traffic is descending, the pilot climbs (Merwin, Wickens, & O'Brien, 1998; Wickens, Helleberg, & Xu, 1999). Thus, descending traffic will induce a climbing maneuver, and the resulting loss of airspeed will reduce the maneuverability of the aircraft to avoid the traffic if additional maneuvers are required. This would account for the greater time spent in both predicted and actual conflict shown when traffic was descending (and the predictor was incorrect), relative to conditions in which traffic was climbing (and the pilot would choose a descending-avoidance maneuver).

One view of the problems pilots encountered when the predictor was in error and the task was particularly difficult is that this represents an undesirable and avoidable complacency effect, reflecting an overtrust in and overreliance upon the automation-based predictor (Parasuraman, Molloy, & Singh, 1993; Parasuraman & Riley, 1997). Although this view is consistent with the modest overestimate of predictor reliability given in the subjective trust measure, we chose instead to characterize this behavior as a more appropriately, and indeed more optimally calibrated, allocation of attention between the automation-based predictor and the raw data (true aircraft position). This allocation policy may be described in much the same way that probability matching in signal detection theory can be shown to be optimal, even if it will lead to some percentage of incorrect outcomes. Pilots always do use the predictor even knowing its unreliability because it is right most of the time, and when correct it is quite helpful for conflict avoidance as Experiment 1 shows.

A complacency effect, in contrast, is indicative of overtrust, a trust that could be expected to be recalibrated once the nature or existence of the unreliability is discovered as an effect we have observed elsewhere (Davison & Wickens, 1999; Merlo et al., 1999). Yet in the current experiment, little or no evidence was found for such recalibration in trust as implicitly measured by raw data monitoring and secondary task performance, contrasted before and after a failure. That is, had such recalibration occurred and pilots become more attentive to the more difficult to process raw data (the traffic aircraft symbol itself) we would have expected to observe a withdrawal of attention from the FFOV task. This withdrawal was not observed as inferred from the FFOV measures. The estimates of failure rate provided further evidence that pilots were reasonably

accurately calibrated in their trust of the system, which corresponded fairly closely with the actual rate (0.133 vs. 0.167, respectively).

Speculating what factors might have caused more severe departures from the optimality than relatively high calibration observed here is important. Results from other experiments suggest the joint influence of the salience and ease of processing and interpretation of the automation attention guidance, and the high difficulty in processing the raw data as leading to a less calibrated overreliance on the automation guidance. Thus, for example, Merlo et al. (1999) found that less salient raw data was hurt more in its detectability by imprecision in an automation attention guidance cue than was more salient raw data. Similarly, Palmer and Degani (1991) found that pilots tended more to follow the incorrect guidance of an automated checklist when the error item was not salient in the raw data. Wickens, Conejo, and Gempler (1999) found continued reliance on unreliable guidance from an automatic target recognition device, but in a paradigm in which true targets were very difficult to distinguish from the false ones, indicated by the unreliable automation. Correspondingly, other data suggests that making the raw data more salient or easier to process can reduce the complacency effect. Molloy and Parasuraman (1994) found that the degree of complacency in automation was reduced when a more intuitive ecological display was used to present the raw data in an engine-monitoring task.

In this study, a feature that may have supported the better integration of raw data (the aircraft symbol) and automation (the prediction symbol), is the object-like integration of the two information sources connected by a line. Such display integration is known to support the cognitive or mental integration imposed by the task better (Wickens & Andre, 1990; Wickens & Carswell, 1995).

Finally, given the relatively high degree of calibration between actual system reliability and trust observed in the current data (and, we assume, the consequent calibration of attention allocation between the predictor and the raw data), we are not altogether surprised that the W display, designed to encourage such calibration, was not effective. There was simply little room for improvement in calibration when the line predictor was employed. However, the W apparently had a few unanticipated effects. It appeared to induce a greater amount of lateral maneuvering, and perhaps because of this increase, coupled with the fact that lateral maneuvers are of greater complexity (higher order) and hence more difficult to implement (Wickens, 1999), the W led to higher control activity. The W display was also associated with slightly greater time (1.9 sec) spent in actual conflict and with improved monitoring of the FFOV task, yielding 0.43 sec faster responses to the events. This fact also may have been responsible for the lowered success of the W display augmentation in avoiding conflicts. That is, any benefits it proffered were realized in more available visual resources to the secondary task, rather than in improved primary-task performance. The extent to which one or both of these effects were the consequence of the W display cannot be determined with certainty.

GENERAL DISCUSSION

The two experiments reported here have demonstrated the benefits and costs of predictive automation. Experiment 1 clearly revealed the separate benefits of both ownship and traffic prediction in terms of both improved performance (flight safety) and reduced workload as measured by the subjective rating of mental demand. The latter workload reduction was not paralleled by a corresponding improvement in secondary-task performance (the FFOV monitoring task), a fact that can possibly be attributed to the added visual requirements of the TV predictor on the traf-

fic display. We should note, however, that no loss in secondary task performance resulted from the added predictor elements and from the resulting increase in visual complexity of the primary-task display.

As revealed in Experiment 2, the attention allocated to the predictor, improving performance when it was correct, produced an anticipated cost when the predictor was in error. This cost was inferred to result from the proportionately greater allocation of attention to the predictor than to the raw data depicting the actual aircraft movement, particularly when the data suggested a more difficult and complex 3-axis conflict. Such an allocation policy can be inferred to represent reasonably optimal behavior for a well-calibrated operator and parallels the policy adopted by pilots in a CDTI study when visual scanning was measured (Ellis & Stark, 1986). As Posner (1978) has pointed out, the allocation of attention in a probabilistic world must have its costs as well as its benefits. The current data however suggest that these costs were not excessive. Pilots never allowed their aircraft to collide with the simulated traffic, and the added time spent in actual conflict (about 10 sec for the descending traffic) could be anticipated as a necessary cost imposed by the more frequent benefits of attending to the predictor.

The current results were of course constrained in their generalizations to aircraft equipped with traffic displays in a free-flight scenario because of the relatively low fidelity of the flight simulation employed here. Nevertheless, the fact that similar workload effects on outside scanning and automation reliability influences on performance have been observed in higher fidelity simulations (Wickens, Helleberg, & Xu, 2000) and that the inferred distribution of attention matches measured scanning strategies (Ellis & Stark, 1986) supports the generalizability of the current research.

In conclusion, it seems that what is important in automation systems that guide the user's attention is not that their reliability be perfect but rather that they are sufficiently reliable to provide more help than bother, that their workload is not excessive, and that operators who use them can deploy attention probabilistically in such a way that performance and safety are not seriously jeopardized when the attention guidance is incorrect.

ACKNOWLEDGMENTS

We acknowledge the support of a grant from NASA Ames Research Center for Experiment 1 (NASA NAG 2–996). Vernol Battiste was the scientific and technical monitor. Experiment 2 was also supported by a grant from NASA Ames Research Center (NASA NAG 2–886). Dr. David Foyle was the scientific and technical monitor.

We thank Ron Carbonari for his invaluable contributions in developing the software for the experiments.

REFERENCES

Andre, A. D., & Cutler, H. A. (1998). Displaying uncertainty in advanced navigation systems. *Proceedings of the 42nd Annual Meeting of the Human Factors & Ergonomics Society,* 31–35.

Barhydt, R., & Hansman, R. J. (1997). Experimental studies of intent information on cockpit traffic displays. In R. Jensen & L. Rakovan (Eds.), *Proceedings of the 9th International Symposium on Aviation Psychology* (pp. 261–267). Columbus: Ohio State University.

Davison, H., & Wickens, C. D. (1999). *Rotorcraft hazard cueing: The effects on attention and trust* (Tech. Rep. No. ARL–99–5/NASA–99–1). Savoy: University of Illinois Institute of Aviation, Aviation Research Lab.

Ellis, S., McGreevy, K., & Hitchcock, L. (1987). Perspective traffic display format and airline pilot traffic avoidance. *Human Factors, 29,* 371–382.

Ellis, S., & Stark, L. (1986). Statistical dependency in visual scanning. *Human Factors, 28,* 421–438.

Gabriel, R. F. (1993). Cockpit automation. *Human factors for flight deck certification personnel final report* (DOT–FAA–RD–93/5—DOT–VNTSC–FAA–93–4, pp. 209–241). Washington, DC: Federal Aviation Administration, Research and Development Service.

Gempler, K. S., & Wickens, C. D. (1998). *Display of predictor reliability on a cockpit display of traffic information* (Final Tech. Rep. No. ARL–98–6/ROCKWELL–98–1). Savoy: University of Illinois, Institute of Aviation, Aviation Research Lab.

Hart, S., & Loomis, L. (1980). Evaluation of the potential format and content of a cockpit display of traffic information. *Human Factors, 22,* 591–604.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA–TLS (task load index): Results of empirical and theoretical results. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload.* Amsterdam: North Holland.

Jensen, R. S. (1981). Prediction and quickening in perspective flight displays for curved landing approaches. *Human Factors, 23,* 355–363.

Johnson, W. W., Battiste, V., & Bochow, S. (1999). A cockpit display designed to enable limited flight deck separation responsibility. *Proceedings 1999 World Aviation Conference,* American Institute of Aeronautics and Astronautics.

Johnson, W. W., Battiste, V., Delzell, S., Holland, S., Belcher, S., & Jordon, K. (1997). Development and demonstration of a prototype free flight cockpit display of traffic information. *Proceedings of the 1997 SAE/AIAA World Aviation Conference,* American Institute of Aeronautics and Astronautics.

Kantowitz, B. H., Hanowski, R. J., & Kantowitz, S. C. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. *Human Factors, 39,* 164–176.

Kirschenbaum, S. S., & Arruda, J. E. (1994). Effects of graphic and verbal probability information on command decision making. *Human Factors, 22,* 406–418.

Kreifeldt, J. G. (1980). Cockpit displayed traffic information and distributed management in air traffic control. *Human Factors, 22,* 671–691.

Lamoureux, T. (1999). The influence of aircraft proximity data on the subjective mental workload of the air traffic controller. *Ergonomics, 42,* 1482–1491.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35,* 1243–1270.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Direction in Psychological Science, 5,* 161–171.

Merlo, J. L., Wickens, C. D., & Yeh, M. (1999). Effect of reliability on cue effectiveness and display signaling (Tech. Rep. No. ARL–99–4/FED–LAB–99–3). Savoy: University of Illinois, Institute of Aviation, Aviation Research Lab.

Merwin, D., O'Brien J. V., & Wickens, C. D. (1997). Perspective and coplanar representation of air traffic: Implications for conflict and weather avoidance. *Proceedings of the 9th International Symposium on Aviation Psychology,* 362–367.

Merwin, D. H., & Wickens, C. D. (1996). *Evaluation of perspective and coplanar cockpit displays of traffic information to support hazard awareness in free flight* (Tech. Rep. No. ARL–96–5/NASA–96–1). Savoy: University of Illinois, Institute of Aviation, Aviation Research Lab.

Merwin, D. H., Wickens, C. D., & O'Brien, J. V. (1998). Display-format-induced biases in air traffic avoidance behavior. *Proceedings of the World Aviation Congress* (98WAC–71). Warrendale, PA: Society of Automotive Engineers.

Molloy, R., & Parasuraman, R. (1994). Automation-induced monitoring inefficiency: The role of display integration and redundant color coding. In M. Mouloua & R. Parasuraman (Eds.), *Human performance in automated systems: Current research and trends* (pp. 224–228). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Moray, N. (1986). Monitoring behavior and supervisory control. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and performance* (Vol. 2, pp. 40-1–40-51). New York: Wiley.

Morphew, E. M., & Wickens, C. D. (1998). Pilot performance and workload using traffic displays to support free flight. *Proceedings of the 42nd Annual Meeting of the Human Factors & Ergonomics Society,* 52–56.

Palmer, E., & Degani, A. (1991). Electronic checklists: Evaluation of two levels of automation. *Proceedings of the 6th International Symposium on Aviation Psychology,* 178–183.

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

Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors, 38,* 665–679.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39,* 230–253.

Planzer, N., & Jenny, M. T. (1995, January–March). Managing the evolution to free flight. *Journal of ATC,* 18–20.

Posner, M. I. (1978). *Chronometric explorations of the mind.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

VanBreda, L. (1999). *Anticipatory behaviour in supervisory control.* Delft, The Netherlands: Delft University Press.

Veldhuyzen, W., & Stassen, H. G. (1977). The internal model concept: An application to modeling human control of large ships. *Human Factors, 19,* 367–380.

Ververs, P. M., & Wickens, C. D. (2000). Designing head-up displays (HUDs) to support flight path guidance while minimizing effects of cognitive tunneling. *Proceedings of the IEA2000/HFES2000 Congress.* Santa Monica, CA: Human Factors and Ergonomics Society.

Vicente, K., & Rasmussen, J. (1992). Ecological interface design. Theoretical foundations. *IEEE Transactions on Systems, Man, and Cybernetics, 22,* 589–606.

Wickens, C. D. (1986). The effects of control dynamics on performance. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (Vol. 2, pp. 39.1–39.60). New York: Wiley.

Wickens, C. D. (1991). Processing resources and attention. In D. Damos (Ed.), *Multiple-task performance* (pp. 3–34). London: Taylor & Francis.

Wickens, C. D. (1998, October). Common sense statistics. *Ergonomics in Design,* 18–22.

Wickens, C. D. (1999). Cognitive factors in aviation. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of applied cognition* (pp. 247–282). New York: Wiley.

Wickens, C. D., & Andre, A. D. (1990). Proximity, compatibility, and information display: Effects of color, space, and objectness of information integration. *Human Factors, 32,* 61–77.

Wickens, C. D., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors, 37,* 473–494.

Wickens, C. D., Conejo, R., & Gempler, K. (1999). Unreliable automated attention cueing for air–ground targeting and traffic maneuvering. *Proceedings of the 43rd Annual Meeting of the Human Factors & Ergonomics Society.*

Wickens, C. D., Helleberg, J., & Xu, X. (1999). Maneuver choice in free flight. *Proceedings of the World Aviation Congress* (1999–01–5591). Warrendale, PA: Society of Automotive Engineers.

Wickens, C. D., Helleberg, J., & Xu, X. (2000). Decision and workload implications of free flight and the cockpit display of traffic information (CDTI). *Proceedings of the IEA 2000/HFES 2000 Congress.* Santa Monica, CA: Human Factors and Ergonomics Society.

Wickens, C. D., & Hollands, J. (2000). *Engineering psychology and human performance* (3rd ed.). New York: Prentice Hall.

Wickens, C. D., Mavor, A. S., & McGee, J. P. (1997). *Flight to the future: Human factors in air traffic control.* Washington, DC: National Academy Press.

Wickens, C. D., Mavor, A. S., Parasuraman, R., & McGee, P. (1998). Airspace system integration: The concept of free flight. In *The future of air traffic control.* Washington DC: National Academy Press.