Identifying Black Swans in NextGen: Predicting Human Performance in Off-Nominal Conditions

Christopher D. Wickens, Alion Science Corporation, Boulder, Colorado, **Becky L. Hooey** and **Brian F. Gore,** San Jose State University Research Foundation at NASA Ames Research Center, Moffett Field, California, and **Angelia Sebok and Corey S. Koenicke,** Alion Science Corporation, Boulder, Colorado

Objective: The objective is to validate a computational model of visual attention against empirical data—derived from a meta-analysis—of pilots' failure to notice safety-critical unexpected events. **Background:** Many aircraft accidents have resulted, in part, because of failure to notice nonsalient unexpected events outside of foveal vision, illustrating the phenomenon of change blindness. A model of visual noticing, N-SEEV (noticing–salience, expectancy, effort, and value), was developed to predict these failures. **Method:** First, 25 studies that reported objective data on miss rate for unexpected events in high-fidelity cockpit simulations were identified, and their miss rate data pooled across five variables (phase of flight, event expectancy, event location, presence of a head-up display, and presence of a highway-in-the-sky display). Second, the parameters of the N-SEEV model were tailored to mimic these dichotomies. **Results:** The N-SEEV model output predicted variance in the obtained miss rate ($r = .73$). The individual miss rates of all six dichotomous conditions were predicted within 14%, and four of these were predicted within 7%. **Conclusion:** The N-SEEV model, developed on the basis of an independent data set, was able to successfully predict variance in this safety-critical measure of pilot response to abnormal circumstances, as collected from the literature. **Applications:** As new technology and procedures are envisioned for the future airspace, it is important to predict if these may compromise safety in terms of pilots' failing to notice unexpected events. Computational models such as N-SEEV support cost-effective means of making such predictions.

INTRODUCTION

In the face of challenges to the future airspace system, brought about by increased passenger travel demand and enduring weather delays, a program of research and development has been initiated for the next generation of the airspace, titled NextGen (Joint Program and Development Office, 2008). This program includes defining a set of new operating procedures to integrate the flight deck with air traffic management, as supported by various technologies, automation tools, and decision aids. Although the increased airspace productivity fostered by these technologies and procedures is being modeled and researched, the negative safety implications when unexpected and unpredicted circumstances prevail are less understood. The objective of the research reported here is to provide a validated computational model of pilots' responses to unexpected events, which in turn can be incorporated into overall pilot performance models (Foyle & Hooey, 2008; Gore, 2008) to evaluate both productivity and safety of NextGen technology and procedures.

The psychology of human response to unexpected events can be approached from two overlapping perspectives. On one hand, ample data exist to show that people's response to unexpected events slows in inverse proportion to

Address correspondence to Christopher D. Wickens, Alion Science Corporation, Micro Analysis and Design, 4949 Pearl East Circle, Suite 300, Boulder, CO 80301; cwickens@alionscience.com. *HUMAN FACTORS,* Vol. 51, No. 5, October 2009, pp. 638-651. DOI: 10.1177/0018720809349709.

event probability, a finding well incorporated in the Hick-Hyman law of response time (Fitts & Posner, 1967; Wickens & Hollands, 2000). On the other hand, one can analyze the three information-processing operations that typically take place in real-world contexts when unexpected events occur: noticing, diagnosing, and responding. Although the processing of all of these may be delayed by low expectancy, more significant is the fact that the first operation may fail altogether: People often do not notice unexpected events, even if these events are relatively salient. This phenomenon is known as *change blindness* (Durlach, 2004; Rensink, 2002; Simons & Levin, 1997; Stelzer & Wickens, 2006; Varakin, Levin, & Fidler, 2004).

In a classic study of situation awareness (SA) breakdowns in aviation, Jones and Endsley (1996) observed that the majority of such breakdowns occurred at the first phase of SA (noticing and perception) rather than at later phases of diagnosis and prediction. Furthermore, tragedies in aviation can be associated with failures to notice critical off-nominal events, such as the failure of a position broadcast in a midair collision over Brazil in 2006 (Command of Aeronautics, 2006; National Transportation Safety Board [NTSB], 2007) or the unintentional decoupling of an autopilot and subsequent low altitude alert in a commercial airline crash into the Everglades (Wiener, 1977). There is an important distinction to be drawn here between "somewhat surprising" unexpected events (which are often responded to more slowly than expected events) and truly surprising ones (which may be missed altogether). Taleb (2007) has referred to these as "gray swans" and "black swans," respectively.

A handful of studies in the aviation psychology literature have examined black swans. The pioneering work of Fischer, Haines, and Price (1980) revealed that pilots were more likely to miss detecting an unexpected runway incursion while flying with a head-up display (HUD) than without, although there were no inferential statistics applied to these data. A subsequent meta-analysis of HUD studies carried out by Fadden, Wickens, and Ververs (2000) confirmed the statistical reliability of the HUD costs on detecting such very rare events, even as the HUD overall supported benefits on most tasks. A series of subsequent studies carried out at National Aeronautics and Space Administration (NASA) Ames, NASA Langley, and University of Illinois examined pilots' detection of a variety of off-nominal events obstructing or endangering their flight path (described later). These studies revealed that detection failures of these events were of a sufficiently high frequency to be of concern and that the number of these failures were elevated when attention was attracted to compelling 3-D displays within the cockpit (Wickens & Alexander, 2009).

The modeling of pilot response delay (or nonresponse) to unexpected events is particularly useful for projections of NextGen procedural safety because modeling can be done relatively quickly and inexpensively, compared with the time and money required to carry out pilot-in-the-loop (PIL) simulations. Modeling can also be very effective for defining and evaluating conceptual systems and procedures for which pilots may not have experience. Hence the subject population for PIL simulations will not be typical of the future population anticipated to execute those procedures. Valid computational models, such as NASA's Man-Machine Integration Design and Analysis System (MIDAS; Gore, 2008), that can make predictions about performance in operationally meaningful units (e.g., seconds delayed, events missed) can fill this gap. Although such models may not be able to offer precise predictions of optimal configurations, they often can identify poor designs or safety-compromising procedures and can be used to narrow the parameter space that should be examined more thoroughly with PIL research.

Our approach to the study of the response to black swans in aviation described here consists of three phases: (a) identifying, through a parameter meta-analysis, pilot response parameters (noticing time and miss rate [MR]) for unusual events; (b) developing and refining a computational model (noticing–salience, expectancy, effort, and value [N-SEEV]) to predict parameters for noticing unexpected events; and (c) validating model predictions against the meta-analysis data (see Gore et al., 2009, for full details).

METHOD: PARAMETER META-ANALYSIS

The study inclusion criteria was driven in a top-down fashion, by the consideration of likely scenarios and technologies to be encountered in NextGen operations, and in a bottom-up fashion, by the consideration of the range of scenarios and technologies studied in the available literature. The aviation human factors literature was thoroughly reviewed from the following sources:

- Annual proceedings on Manual Control
- Digital Avionics System Proceedings
- *IEEE Transactions of Systems, Man, Cybernetics, Part A: Systems and Humans*
- *IEEE Transactions of Systems, Man, Cybernetics, Part C: Applications and Reviews*
- *International Journal of Aviation Psychology*
- *International Symposium on Aviation Psychology*
- *Human Factors*
- *NASA Technical Reports Server*
- *Proceedings of the Human Factors and Ergonomic Society Annual Meeting*
- USA/Europe Air Traffic Management R&D Seminar (http://www.atmseminar.org/)

Studies were included in the meta-analysis if they met the following criteria:

1. Were carried out in a reasonable or high-fidelity aviation simulation environment
2. Presented some unexpected event, such as an engine failure or runway incursion
3. Were sufficiently descriptive of this event, so the location of visual evidence for the event and the pilot's level of expectancy could be ascertained with reasonable certainty
4. Presented performance data on the mean time to detect, or the proportion of times the event was noticed

A total of 34 studies that met the above criteria were identified. Within these, 25 studies included data that could be uniquely categorized in terms of five categorical variables: (a) whether the event occurred during taxi, departure, cruise, or approach; (b) whether an HUD was present or not; (c) whether the off-nominal event was visible out the window (OTW) or head down in the cockpit; (d) whether a highway-in-the-sky (HITS) display was present or not; and (e) whether the event was truly surprising (e.g., the last landing of the experiment within an otherwise failure-free series of landings; a black swan) or simply "unexpected" (e.g., a failure of one system late in an experiment, but following an earlier failure of a different system; a gray swan). (The remaining studies were used for analyses of other variables not reported here, but see Hooey et al., 2009.) Table 1 lists the studies included. The first column contains a letter identification by which the study can be located in the reference list. The remaining columns indicate whether the study contributed data to each of the five variables. Note that a study might not contribute at all to a particular analysis (for example, if data were not reported separately for phase of flight in the original paper or article, they were excluded from phase-of-flight analyses here) but is included in the other analyses. Also, a single study may contribute to both levels of a variable if the original study compared both levels (e.g., HUD vs. no HUD).

Our initial intent in this meta-analysis was to include data from NTSB accident reports. However, careful examination of the most relevant of these revealed that in nearly all cases, pilots eventually detected the unusual event (very rapidly if it affected handling qualities), so MR was not a viable variable, and noticing time was very difficult to extract from the flight data recording information provided without the benefit of video or eye-tracking equipment, which was typically available in the simulation studies. Across studies, off-nominal events included items such as an aircraft or radio tower in or near the line of travel (visible in the forward view), a runway offset, a runway incursion, or a warning light or severe weather alert on the cockpit instrument panel.

RESULTS: PARAMETER META-ANALYSIS

We first examined the 64 cells formed by the $4 \times 2 \times 2 \times 2 \times 2$ "design" of the five factors listed earlier (phase of flight, HUD use, event location, HITS use, event expectancy) and found that several of them were unpopulated by any valid experimental data or had too few observations to

TABLE 1: Studies Included in Meta-Analysis

	Phase of Flight				Expectancy		Location		HUD		HITS	
Study*	Depart	Cruise	Approach	Taxi	Black Swan	Gray Swan	OTW	Cockpit	HUD Yes	HUD No	HITS Yes	HITS No
A1			X				X			X	X	
A2			X				X			X	X	
B1			X		X		X			X	X	
B2			X		X		X			X	X	
C				X	X			X	X	X		X
D	X		X			X	X			X		X
E			X		X	X	X					X
F			X	X	X	X	X	X	X	X		X
G			X				X			X		X
H						X	X		X	X		X
I			X	X	X	X		X	X	X		X
J1			X		X		X			X	X	
J2			X		X		X			X		X
K1		X			X		X			X	X	
K2		X			X		X			X		X
L				X		X	X			X		X
M		X				X		X		X		X
N						X		X				
O			X		X		X		X		X	
P			X		X		X		X			X
Q1			X				X			X	X	
Q2			X				X			X		X
Q3			X				X			X	X	
Q4			X					X		X	X	
R		X			X		X			X		X

Note. Each letter represents a unique publication (see code in reference list for full citation). In the event that a publication reported multiple studies, separate studies are delineated by number. HUD = head-up display; HITS = highway-in-the-sky display; OTW = out the window.

contribute to reliable estimates of mean response time (RT) or event detection rate. Very few of the studies reported RT, so this measure was not considered powerful enough to draw valid statistical conclusions. Because small sample size existed in several of the rows, columns, or cells of the five-factor design, considerable pooling of data across these dimensions was required, as described later (and see Gore et al., 2009, for details). Because the different studies that contributed to each cell of the design often varied greatly in their sample size, we weighted their contribution proportional to sample size (i.e., statistical power), a procedure often performed in meta-analyses. We accomplished this weighting by simply summing the two terms of the ratio, number of events detected and total number of events experienced, across all studies within a cell of the relevant comparison.

The pooling procedures eventually yielded five categorical contrasts producing highly reliable statistical effects on detection performance, which was expressed as MR. These five were not intended to represent an exhaustive examination of all effects in the data, as would be the objective of a traditional experimental design, with control exerted on sample size for all cells. Rather, they were intended to balance three criteria: (a) assess five operationally important effects that were observed, (b) have sufficient statistical power to reveal reliable effects, and (c) provide three contrasts (3, 4, and 5, described later) whose parameters could be well captured by the N-SEEV model and hence serve as a target for validation.

Chi-square tests were used to assess whether the relative frequency count of missed versus nonmissed events was statistically equivalent across the level of another variable. Subsequently, where appropriate, further chi-square tests with a log-linear analysis were conducted to determine whether a difference observed might itself be modulated by a second factor. A liberal alpha level of .1 was adopted for all analyses. Given the relatively small number of studies available, and the exploratory nature of this meta-analysis, it was felt that this was an appropriate trade-off of Type I and Type II errors.

The chi-square approach was akin to extracting a single MR from each study and subjecting these data to an ANOVA. However, the ANOVA treats studies with a high sample size (and hence a reliable estimate of MR) as equivalent to those with a very low sample size (an unreliable estimate). As a consequence, the high variability of the low-sample-size studies would often contribute a great deal of variance to the data, sometimes creating highly nonnormal distributions that grossly violated ANOVA assumptions. A second problem with the ANOVA approach is that certain cells that were to be compared were populated by only one or two studies, thus creating a very low sample size, which further constrained statistical power. The chi-square approach that we adopted using pooled MRs increased the sample size (the denominator) and hence statistical power relative to the ANOVA.

Phase of Flight

An analysis of MR (that is, the rate at which pilots failed to detect an off-nominal event) revealed that across all 25 studies in our analysis (pooled across all other variables), the probability of missing an off-nominal event was highest during departures (MR = 0.50), followed by cruise (MR = 0.47), arrival or approach (MR = 0.39), and taxi (MR = 0.20), $\chi^2(3) = 34.61$, $p < .001$. The reader is cautioned in interpreting the departure MR, however, as this was composed of only one study with eight pilots. These MRs may reflect an expectancy effect, as pilots tend to be more vigilant and aware of both the traffic environment and their aircraft status in the terminal area, making event detection during the arrival and taxi phases more likely than in the cruise and departure phases.

The HUD Effect

MR data used to examine the effects of HUD as a function of expectancy and event location are shown in Table 2, pooled across phase of flight and HITS use. (There were no studies that included both a HITS and a HUD.) Low expectancy refers to first failure trials (black swan), and higher expectancy refers to the data from all other trials (gray swans). The results revealed an overall detection cost for flying with a HUD, relative to a head-down display, $\chi^2 = 4.13$, $p < .05$. A significant finding of nonindependence between HUD use and event

TABLE 2: Meta-Analysis Miss Rates of HUD by Expectancy and HUD by Event Location

	Expectancy		Event Location	
	Low Expectancy (Black Swan)	Higher Expectancy (Gray Swan)	OTW	Cockpit
HUD	.37	.40	.36	.46
No HUD	.48	.28	.27	.51

Note. HUD = head-up display; OTW = out the window.

TABLE 3: Meta-Analysis Miss Rates of Expectancy by Event Location

	Expectancy		
Location	Low (Black Swan)	Higher (Gray Swan)	Mean
OTW	.50	.23	.29
Down	.41	.41	.39
Mean	.48	.29	

Note. OTW = out the window.

expectancy, $\chi^2 = 5.93$, $p < .02$, pooling across event location (left side, Table 2) revealed that this HUD cost was amplified in both its magnitude and its statistical significance when the event was the higher-expectancy gray swan event, $\chi^2 = 5.59$, $p < .05$, relative to the black swan event, $\chi^2 = 1.87$, $p = .17$. Indeed, in the latter case, this nonsignificant effect is in the opposite direction, an effect that may reflect the fact that the no-HUD condition contained several studies with a HITS display, demanding more visual attention head down instead of OTW, where many of the off-nominal events were located. The influence of the HITS in inducing head-down scanning is discussed later.

The HUD data were also analyzed by event location, pooled across expectancy (right side, Table 2). Note that no events were located on the HUD itself. These data reveal a significant HUD cost for outside events (e.g., a runway incursion), $\chi^2 = 4.63$, $p < .05$; however, there is no significant difference with or without the HUD ($p > .10$) for events located within the cockpit. The former results reflect the classic Fischer et al. (1980) effect, whereby HUDs were found to obscure detection of unexpected OTW events.

The Event Location Effect

The data used to examine the effect of event location as a function of event expectancy were pooled across flight phase, HUD use, and HITS use and are shown in Table 3.

A chi-square analysis revealed a main effect of location, with the mean MR for cockpit events (MR = 0.39) higher than those for OTW events (MR = 0.29), $\chi^2 = 9.88$, $p < .01$. Because data from some studies were not reported as a function of both expectancy and location, the mean values may include data that were not used in an interaction analysis, an exclusion that accounts for the discrepancy between the mean of the two levels of each variable and the grand mean shown in the table. Importantly, this main effect was moderated by a strong interaction, $\chi^2 = 8.05$, $p < .01$. When event expectancy was low (black swan; first failures, left column), there was a nonsignificant OTW cost ($p > .10$). But when the event expectancy was higher (gray swan; subsequent failures, right column), there was a significant benefit if the event was located in the forward outside view, relative to in the cockpit, $\chi^2 = 22.35$, $p < .01$. Certainly this significant benefit is consistent with the pilots' general OTW vigilance (and procedures recommended

by the Federal Aviation Administration to keep eyes out more than half of the time; U.S. Department of Transportation, 2000). Why this benefit was not significant for first failure trials remains unclear. Importantly, however, the 0.50 value for the OTW black swan cell may be elevated somewhat because this condition in particular contains a preponderance of studies with a HITS display, drawing attention downward (see 5, later).

The Event Expectancy Effect

Using the data in Table 3, we compared across the two columns to examine the influence of truly surprising black swan events (here, always the first, or only, off-nominal event) and merely unexpected gray swan events, which were events subsequent to the first failure. As is evident from the table, expectancy had no effect on detection of head-down events located in the cockpit (bottom row). But for OTW events (top row), there was a large, significant cost for the totally unexpected black swan events (MR = 0.50) compared with the subsequent gray swan events (MR = 0.23), $\chi^2 = 24.7$. The low expectancy cost for OTW events is certainly predictable. The absence of such a cost for events within the cockpit was, to us, somewhat surprising. We infer that items related to confounding or offsetting effects of other variables, differing between the two means in the middle row of Table 3, also differed between the sets of studies compared.

HITS Effect

To examine the HITS effect, we compared the presence or absence of a HITS display only when the event was both unexpected and outside the cockpit. Other conditions were eliminated from the HITS comparison for three reasons: (a) because there were too few data points available (e.g., high expectancy trials) or because it would not make sense to pool the data (including both OTW and cockpit events); (b) because inclusion of these other conditions would blur the impact of head-down attentional tunneling to the HITS, a phenomenon that has been the focus of considerable research (Wickens & Alexander, 2009); and (c) because we wished to preserve characteristics that best matched the N-SEEV model runs reported later. This analysis, a simple contrast, revealed that the highly surprising OTW events were missed with far greater frequency when flying with the HITS (MR = 0.55) than without (MR = 0.33), $\chi^2 = 7.01$, $p < .01$, hence reflecting the well-known HITS-induced attentional tunneling effect (Wickens & Alexander, 2009). As noted earlier, the last three contrasts provided empirical data targets to be predicted by our N-SEEV noticing model, as described next.

METHOD: N-SEEV MODEL IMPLEMENTATION

The N-SEEV model is an elaboration of the SEEV model (Wickens, Goh, Helleberg, Horrey, & Talleur, 2003; Wickens et al., 2008). SEEV predicts how visual attention (saccadic eye movement) is guided in large-scale environments by the *salience* of locations, inhibited by the *effort* required to move attention across the visual workspace, and attracted to locations according to the *expectancy* of seeing an event at a particular location and the *value* of that event (or cost of missing it). In the SEEV model, the value of an area of interest is equal to the priority of the task served by that area multiplied by the relevance of an event at that area to the task in question. A computational version of the SEEV model drives the eyeball around an environment, such as an aircraft cockpit, according to the influence of the four SEEV parameters. Contributions of the four components are additive and can be provided different weightings. Effort is proportional to the separation of displays, and salience is heavily dictated by the contrast between locations and the background, following the salience model of Itti and Koch (2000). For example, the simulated eyeball following the model will fixate more frequently on areas with a high bandwidth (and hence a high expectancy for change) as well as areas that support high-value tasks, such as maintaining stable flight (Wickens et al., 2008).

The N-SEEV model (McCarley, Wickens, Steelman-Allen & Sebok, 2009; Wickens et al., 2009, allows SEEV to drive steady-state scanning but then imposes a to-be-noticed event (TBNE) somewhere in the environment. This event is associated with a salience measure, derived from a computational model of Itti

and Koch (2000) and augmented to include the salience of changes (Resnick, 2002, see Steelman-Allen, McCarley, Wickens, Sebok, & Bzostek, 2009, for details). Each TBNE is also associated with expectancy and value. For example, a red-flashing warning is quite salient and valuable to be noticed but potentially unexpected. A runway incursion, although valuable to be noticed, may be neither expected nor salient.

N-SEEV associates these parameters with numeric values for the TBNE and predicts a noticing time as a function of where the eye is fixated relative to the TBNE. Because the eye scans across the cockpit environment (as driven by SEEV), the model will actually predict a *distribution* of noticing times (e.g., short if close to the TBNE, long if far, as mediated by eccentricity). If the model is run multiple times, this will capture the distribution of eccentricity and hence the distribution of noticing times. This distribution can be interpreted as a cumulative probability function, generating the probability that the location of the TBNE will be fixated within time *T*. Parameters of the model can then be adjusted according to the additional assumption that if the area of the TBNE is not fixated within some criterion time (T_c), then the event will be missed. In this way, the model, if run repeatedly, can generate an MR estimate (see McCarley et al., 2009; Steelman-Allen et al., 2009, for details).

The N-SEEV model was initially validated against two classes of empirical data sets. First, a set of three general aviation (GA) studies (Wickens et al., 2003) along with a pilot visual scanning study involving the use of a Boeing 747 simulator (Sarter, Mumaw, & Wickens, 2007) were used to validate the parameters of SEEV. Through this effort, it was possible to predict more than 90% of the variance of percentage dwell time within different areas of interest in the GA studies and 75% of the variance in percentage dwell within the automated Boeing 747 cockpit (see McCarley et al., 2009; Wickens et al., 2009). Hence, the SEEV component of the N-SEEV model was considered validated.

Second, the noticing component of the N-SEEV model was validated against noticing time and MR data collected by Nikolic, Orr, and Sarter (2004) in an experiment evaluating participants' ability to notice simulated flight mode annunciator changes in a visual environment that varied in its clutter, spatial layout, and event salience, all parameters that could be incorporated into N-SEEV. With repeated iteration of the model, this exercise enabled identification of particular parameter settings that could accurately predict both the noticing time and MR data from the various conditions of the Nikolic et al. experiment. In particular, the assumption of a scan rate of 2 fixations per second and a miss criterion of 7.5 s (i.e., if the target was not fixated within 7.5 s, it would be missed) were found to provide the best fit to the existing data, yielding correlations between predicted and observed values for both noticing time and MR of greater than 0.95 (Wickens et al., 2009).

RESULTS: N-SEEV MODEL META-ANALYSIS VALIDATION

The next step in this effort was to validate the model against the meta-analysis data described earlier, as the empirical data represented a robust data set that was highly representative of a range of actual flight operations. The model was applied to the cockpit layout rendered in Figure 1. Within this figure, six scenarios, composed of three contrasts each with two levels, were chosen from the meta-analysis for validation. These were OTW versus cockpit location for somewhat surprising events, high versus low expectancy detection of OTW events, and presence versus absence of HITS for black swan OTW events. Each scenario was characterized by parameters of N-SEEV. Unless otherwise noted, all TBNE events were simulated in the model by the onset of a gray circle, positioned OTW, of a diameter half the size of the master caution and warning (MCW) box positioned in the top center of Figure 1. Unless otherwise noted also, the attention demand of flying was established by setting relatively high (0.7 on scale of 0 to 1.0) bandwidth and value levels for the head-down attitude direction indicator (ADI), with a lower (0.3) setting for the OTW view. These values enabled the SEEV component of the N-SEEV model to produce the actual scan percentages observed in GA flight simulations (Helleberg & Wickens, 2003; Wickens et al., 2003; Wickens, Helleberg, & Xu, 2002).

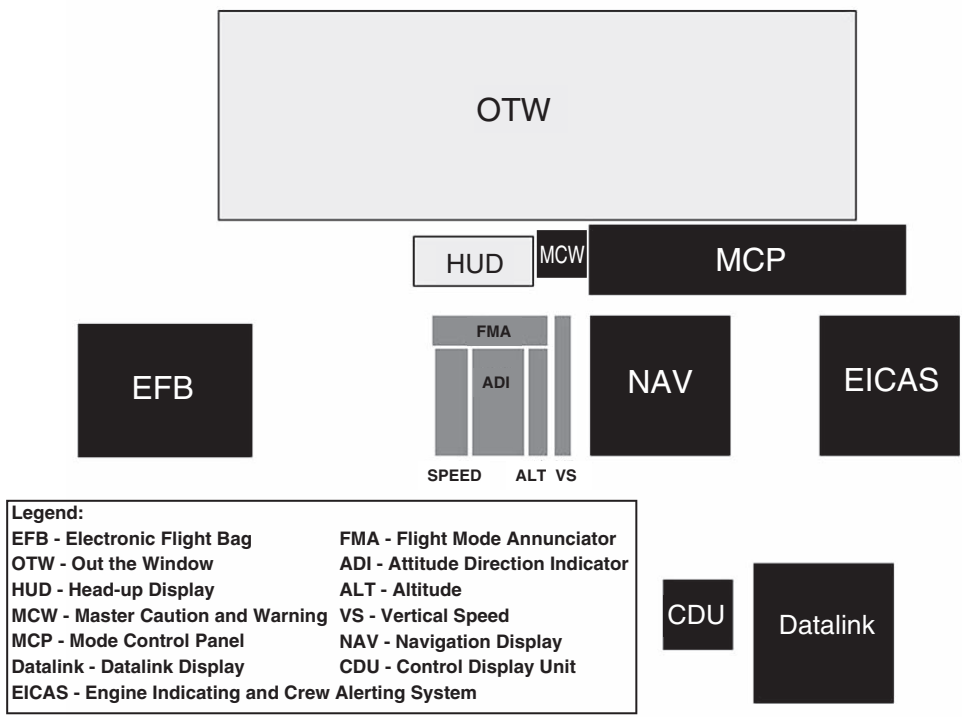

Figure 1. Instrument panel layout upon which model was exercised. The Area of Interest (AOI) corresponding to the to-be-noticed event is not shown here but was positioned as required for the particular model runs.

The three specific contrasts were accomplished by setting parameters as follows:

- The location effect for gray swan (somewhat surprising) events was tested by presenting the TBNE either at the top of the OTW (see Figure 1) or head down, just below the ADI.
- The expectancy effect was tested by varying the bandwidth of the TBNE between 0 (truly surprising black swan) and 0.2 (somewhat surprising gray swan).
- The HITS cost for truly surprising (bandwidth = 0) OTW events was tested by varying the value and bandwidth parameters of the ADI with parameter setting of either [1.0, 1.0] (HITS present) or [0.3, 0.4] (HITS absent). In addition, for the non-HITS condition, a 60% increase in the OTW value parameter (from 0.3 to 0.5) was provided as well as a more modest 20% increase in the value parameter of the altimeter, vertical speed indicator (VSI), and heading indicator (Navigation display). This adjustment was made because the pilot would use these instruments to compensate for the direct lateral and vertical guidance otherwise provided by the HITS (see Gore et al., 2009, for more details and Wickens et al., 2008, for more details on visual scanning with a HITS).

Six model runs were then carried out, with each run iterated 1,000 times to generate the requisite Monte Carlo distribution of noticing times. Using the same N-SEEV model parameters established by the validation work described in the previous section (also see McCarley et al., 2009; Wickens et al., 2009), a set of MR predictions were generated across the six conditions. These are shown on the *x*-axis of Figure 2. The *y*-axis depicts the corresponding obtained MRs from the meta-analysis. Connected pairs of points represent the two conditions compared within each of the three contrasts, as labeled (i.e., location effect, expectancy effect, and HITS effect).

Four general features of this model validation are noteworthy. First, the overall correlation, across the six data points between predicted

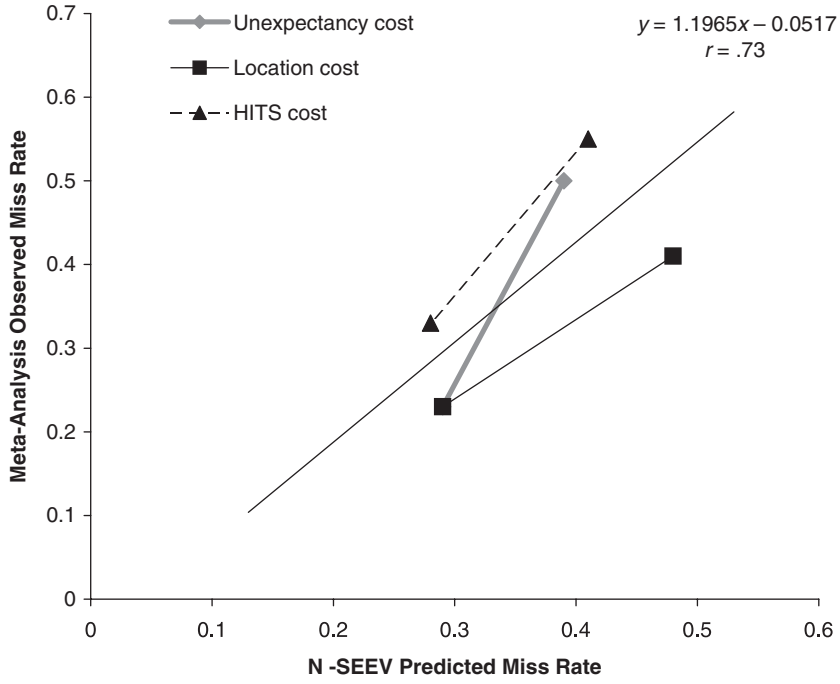

Figure 2. Validation of N-SEEV (noticing–salience, expectancy, effort, and value) model predicted miss rate (*x*-axis) against measured MR from the meta-analysis. Lines connect the two points within each dichotomous comparison.

and obtained MR, was $r = .73$, an adequate fit given the heterogeneity of variables that were varied across the six conditions and the diversity of empirical sources contributing to the meta-analysis. Second, regression of the six points collectively generated a slope value (1.2) reasonably close to 1.0 and an intercept (0.08) reasonably close to zero. This means that not only are changes in model predictions echoed in changes in obtained data (the .73 positive correlation), but the actual values of predicted MRs correspond closely to the actual values obtained. Third, each contrast by itself produced the predicted positive slope, varying from 2.7 (expectancy effect) to 1.7 (location effect) to 0.95 (HITS effect). Fourth, we note that all of the six empirical data points were predicted within 14% (on an absolute scale, that is, for example, 54% observed, 40% predicted). Furthermore, four of the data points were predicted within 7%.

Following this validation, a final phase of the research effort, reported in full detail in Gore et al. (2009), was designed to predict the impact of different procedures and displays projected to be implemented in the next generation of the airspace. For example, we found that the procedure for pilots to monitor a cockpit display of traffic information, supporting responsibility for self-separation from other aircraft, led to a predicted MR of black swan OTW events of 0.52. When this was coupled with the cognitive demands (simulated by restricting field of view or attentional narrowing) of addressing an engine failure, the MR escalated to 0.83. The requirement to monitor a specialized display, supporting very closely spaced parallel approaches, led to an MR of 0.58 for engine failure indicators. It should be noted that these values assumed only a single pilot.

DISCUSSION

We have discussed two important and interrelated issues here: the performance of pilots detecting very unusual events and the ability of a psychologically based computational model to predict such detection. Regarding the first of these, our meta-analyses revealed substantial

performance decrements, with MR averaged across conditions of 32%. On one hand, such a level of performance might well be considered disconcerting for aviation safety. But on the other hand, such misses will occur quite infrequently, because the base rate of these off-nominal black swan events is, by definition, exceedingly low (but not impossible). Furthermore, the results from these high-fidelity flight simulations certainly replicate what is now well known regarding change blindness and inattentional blindness in the real world (Durlach, 2004; Rensink, 2002; Sarter et al., 2007; Simons & Levin, 1997; Stelzer & Wickens, 2006; Wickens & Alexander, 2009; Wickens, Thomas, & Young, 2000). That is, people simply do a poor job of noticing changes (events) when (a) these are unexpected, (b) they are not salient, and (c) they occur outside of foveal vision, all conditions that typified the events analyzed in our meta-analysis.

Regarding the linkage between the meta-analysis and the N-SEEV model fitting, we were gratified that a model, initially developed for scanning (Sarter et al., 2007; Wickens et al., 2003), and the noticing component initially validated with expected events (Nikolic et al., 2004; Wickens et al., 2009) could capture the MR of unexpected events with substantial predictive power (Figure 2) simply by adjusting the bandwidth (expectancy) parameter to 0, a cognitively plausible manipulation (Moray, 2003).

One slight anomaly with the model fit, shown in Figure 2, is that the "expectancy effect" generated by the model underestimates its magnitude in the meta-analysis (the slope function of 2.7). One plausible explanation for this underestimation is that in the actual data from the meta-analysis, a greater number of studies included in the totally surprising black swan condition were flown with a HITS than those in the somewhat surprising gray swan cell, hence elevating the measured cost of low expectancy on the *y*-axis, because the HITS is found to elevate the MR. A second explanation is that the "first failure" feature used to estimate the no-expectancy black swan may be qualitatively, as well as quantitatively, different from variations in expectancy after that first failure has occurred (or after the pilot realizes that it could occur). In contrast, the model expresses a quantitative and linear difference in predicted expectancy by lowering the additive parameter from 0.2 to 0. We note that some models of subjective probability do not make this linear assumption at very low probabilities (e.g., prospect theory of Tversky & Kahneman, 1981). Had similar nonlinearities between low and no expectancy been invoked in the current model, a better fit for this effect would have been obtained.

When translating the model-computed noticing time distribution to an MR prediction, one major assumption must be highlighted and could plausibly be debated. A speed–accuracy trade-off criterion (T_c) was adopted such that if the TBNE was not fixated within time T_c, it would be missed. For the model, we established T_c to be 7.5 s, and there was no firm basis for doing so, other than that such a criterion provided the best fit to both speed and accuracy data from Nikolic et al. (2004; see Gore et al., 2009; McCarley et al., 2009) and provided the best correlation with the MR data from the current meta-analysis.

Indeed, the issue of understanding the speed–accuracy trade-off within detection remains challenging (Wickens & Hollands, 2000); for sometimes it is a trade-off, as it typically is in visual search tasks (Drury, 1994; McCarley et al., 2009), but at other times, it is seemingly a "trade-on" such that longer responses will be associated with more, rather than fewer, errors, and this trade has not been explicitly examined in noticing (as opposed to search) tasks.

We acknowledge that the current research has at least five important limitations. First, in spite of the high fidelity of the studies covered in the meta-analyses, real-world failures (truly naturalistic observations) were not included. Very few data of this sort exist in the literature given the difficulties, and safety issues, associated with exposing pilots to off-nominal event in the actual operating environment.

Second, the six conditions chosen for validation could have been considered nonarbitrary or nonrandom. For example, we specifically focused the HITS validation on studies where the off-nominal event was truly surprising (black swan) and OTW, and the event location validation was restricted to gray swans.

Certainly, had the primary purpose of the research effort been to evaluate the effects of HUD, HITS, and other variables on off-nominal event detection, we would have taken a different, more controlled approach, and indeed, that approach has been taken elsewhere with regard to the HITS (Wickens & Alexander, 2009) and HUDs (Fischer et al., 1980; Weintraub, Haines, & Randle, 1985) and event location (Hofer, Braune, Boucek, & Pfaff, 2001). Instead, here the focus was explicitly on choosing robust, reliable, and operationally meaningful contrasts that could be used to evaluate the model. More specifically, the resulting contrasts selected were powerful and highly significant, thereby presenting a greater challenge for model prediction. Furthermore, they were selected on the basis of conditions in which a substantial *N* existed in the meta-analysis, hence providing considerable power and confidence that the meta-analytic effects were significant (and hence "real" targets of prediction).

A third concern relates to the absence of perfect correspondence between the parameters selected for N-SEEV and the conditions of the meta-analysis studies included within a category. The problem is that in some cases, the different studies within a category may have differed (it was not always determinable from the published reports) on some parameter in the model other than that which was varied in the contrast. Here, parameters were estimated in a way that was most representative of all the studies, as best as possible.

A fourth concern is that the conditions contrasted in the meta-analysis were not entirely free of confounding variables. As we have noted, the effect of HUD presence versus absence was somewhat confounded because the no-HUD studies contained some studies with a HITS, whereas none of the HUD studies had a HITS. And as we have noted, the expectancy contrast shown in Table 3 and Figure 2 was also partially confounded with the presence or absence of a HITS. This confound did not, however, affect the location effect or the HITS contrast itself.

Finally, one can always question the extent to which the off-nominal events chosen from the meta-analysis, based of course on experimental data, were representative of real-world effects. Such is always the case with data collected in laboratory simulations, no matter how high is the fidelity. To this, we can point only to the very real existence of pilots missing black swan events in real-world mishaps to establish that there is at least a modest degree of representativeness.

In conclusion, we are optimistic regarding the value of the N-SEEV model to predict circumstances when future technology and procedures may produce very unexpected events, for which delays or failures of pilot detection can compromise safety. Such prediction may trigger alteration of procedures or targeted PIL simulations to cross-check, and hopefully further validate, the model predictions. We believe that this work illustrates the value of a methodology that focuses on the rich interplay between theory, aggregate data, and models.

ACKNOWLEDGMENTS

This research was supported by a cooperative agreement from National Aeronautics and Space Administration's (NASA) NextGen-Airspace Project (Airspace Super Density Operations, NRA No. NNX08AE87A) to San Jose State University (PI: Brian.F.Gore@nasa.gov). The authors would like to thank NASA's technical monitor (Dr. David Foyle) for his overview of the project and all reviewers for their comments on the present document. The authors acknowledge the invaluable contributions of Jason McCarley and Kelly Steelman-Allen for N-SEEV (noticing–salience, expectancy, effort, and value) model implementation, of Ellen Salud and Shaun Hutchins for work on the meta-analysis, and of Julie Bzostek for model evaluation.

REFERENCES

[A] Alexander, A. L., & Wickens, C. D. (2005). *3D navigation and integrated hazard display in advanced avionics: Performance, situation awareness, and workload* (Tech. Rep. AHFD-05-10/NASA-05-2). Savoy: University of Illinois, Aviation Human Factors Division.

[B] Alexander, A. L., Wickens, C. D., & Hardy, T. J. (2005). Synthetic vision and the primary flight display. *Human Factors*, *47*, 693–707.

[C] Arthur, J. J., Prinzel, L. J., Williams, S. P., & Kramer, L. J. (2004). *Synthetic vision enhanced surface operations and flight procedures rehearsal tools* (NASA/TP-2004-213008). Hampton, VA: NASA.

[D] Arthur, J. J., Prinzel, L. J., Bailey, R. E., Shelton, K. J., Williams, S. P., Kramer, L. J., & Norman, R. M. (2008). *Head-worn display concepts for surface operations for commercial aircraft*. (NASA/TP-2008-215321). Hampton, VA: NASA.

[E] Bailey, R. E., Kramer, L. J., & Prinzel, L. J. (2006). Crew and display concepts evaluation for synthetic enhanced vision systems. In *Proceedings of SPIE, the International Society for Optical Engineering* (Vol. 6226, pp. 62260G.1–62260G). Bellingham, WA: Society of Photo-Optical Instrumentation Engineers.

Command of Aeronautics. (2006). *Final Report A-0-22 CENIPA / 2008*. Brasilia, Brazil: CENIPA (Centro de Investigação e Prevenção de Acidentes Aeronáuticos).

Drury, C. (1994). The speed–accuracy trade-off in industry. *Ergonomics*, *37*, 747–763.

Durlach, P. J. (2004). Change blindness and its implications for complex monitoring and control systems design and operator training. *Human–Computer Interaction*, *19*, 423–451.

Fadden, S., Wickens, C. D., & Ververs, P. M. (2000). Costs and benefits of head-up displays: An attention perspective and a meta analysis. In *2000 World Aviation Congress* (Paper No. 2000-01-5542). Warrendale, PA: Society of Automotive Engineers.

[F] Fischer, E., Haines, R. F., & Price, T. A. (1980). *Cognitive issues in head-up displays* (NASA Technical Paper 1711). Moffett Field, CA: NASA Ames Research Center.

Fitts, P. M., & Posner, M. I. (1967). *Learning and skilled performance in human performance.* Belmont, CA: Brock-Cole.

Foyle, D. C., & Hooey, B. L. (Eds.). (2008). *Human performance modeling in aviation.* Boca Raton, FL: Taylor and Francis/ CRC.

Gore, B. F. (2008). Human performance: Evaluating the cognitive aspects. In V. Duffy (Ed.), *Handbook of digital human modeling* (pp. 32.1-32.18). Boca Raton, FL: Taylor and Francis.

Gore, B. F., Hooey, B. L., Wickens, C. D., Sebok, A., Hutchins, S., Salud, E., Small, R., Koenecke, C., & Bzostek, J. (2009). *Identification of pilot performance parameters for human performance models of off-nominal events in the NextGen environment* (Final Report NRA No. NNX08AE87A). San Jose, CA: San Jose State University.

[G] Helleberg, J. (2005). *Effects of a final approach runway occupancy signal (FAROS) on pilots' flight path tracking, traffic detection, and air traffic control communications*. McLean, VA: MITRE Corporation.

Helleberg, J., & Wickens, C. D. (2003). Effects of data link modality and display redundancy on pilot performance: An attentional perspective. *International Journal of Aviation Psychology*, *13*, 189–210.

[H] Hofer, E. F., Braune, R. J., Boucek, G. P., & Pfaff, T. A. (2001). *Attention switching between near and far domains: An exploratory study of pilots' attention switching with head-up and head-down* (D6-36668). Seattle, WA: Boeing Commercial Airplane Co.

[I] Hooey, B. L., Foyle, D. C., & Andre, A. D. (2000). Integration of cockpit displays for surface operations: The final stage of a human-centered design approach. *SAE Transactions: Journal of Aerospace*, *109*, 1053–1065.

Hooey, B. L., Wickens, C. D., Salud, E., Sebok, A., Hutchins, S., & Gore, B. F. (2009). Predicting the unpredictable: Estimating human performance parameters for off-nominal events. In P. M. Vidulich and P. Tsang, *Proceedings of the 15th International Symposium on Aviation Psychology* [CD-ROM]. Dayton, OH: Wright State University.

[J] Iani, C., & Wickens, C. D. (2007). Factors affecting task management in aviation. *Human Factors*, *49*, 16–24.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506.

[K] Johnson, N. R., Wiegmann, D. A., & Wickens, C. D. (2005). *Effects of advanced cockpit displays on general aviation pilots' decisions to continue visual flight rules (VFR) flight into instrument meteorological conditions (IMC)* (AFHD-05-18/ NASA-05-6). Savoy: University of Illinois, Aviation Human Factors Division.

Joint Program and Development Office. (2008). *Next generation air transportation integrated work plan* (Version 1.0). Washington, DC: Federal Aviation Administration.

Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space, Environmental Medicine*, *67*(6), 507–512.

[L] Lorenz, B., & Biella, M. (2006). Evaluation of onboard taxi guidance support on pilot performance in airport surface navigation. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 111–115). Santa Monica, CA: Human Factors and Ergonomics Society.

McCarley, J., Wickens, C. D., Steelman-Allen, K., & Sebok, A. (2009). *Control of attention: Modeling the effects of stimulus characteristics, task demands, and individual differences.* (NASA Final Report, ROA 2007, NRA NNX07AV97A). San Jose, CA: San Jose State University.

Moray, N. (2003). Monitoring, complacency, scepticism and eutactic behaviour. *International Journal of Industrial Ergonomics*, *31*(3), 175–178.

[M] Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology*, *8*, 47–63.

National Transportation Safety Board. (2007, May 2). [Letter to Marion C. Blakey of the Federal Aviation Administration.] Available from National Transportation Safety Board Web site: http://www.ntsb.gov/Recs/letters/2007/A07_35_37.pdf

Nikolic, M. I., Orr, J. M., & Sarter, N. B. (2004). Why pilots miss the green box: How display context undermines attention capture. *International Journal of Aviation Psychology*, *14*(1), 39–52.

[N] Olson, W. A., & Sarter, N. B. (2001). Management-by-consent in human–machine systems: When and why it breaks down. *Human Factors*, *43*, 255–266.

[O] Prinzel, L. J., Kramer, L. J., Arthur, J. J., Bailey, R. E., & Comstock, R. J. (2004). Comparison of head-up and head-down "highway in the sky" tunnel and guidance concepts for synthetic vision displays. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 11–15). Santa Monica, CA: Human Factors and Ergonomics Society.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology*, *53*, 245–277.

Sarter, N. B., Mumaw, R., & Wickens, C. D. (2007). Pilots monitoring strategies and performance on highly automated glass cockpit aircraft. *Human Factors*, *49*, 347–357.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Science*, *1*, 261–267.

Stelzer, E. M., & Wickens, C. D. (2006). Pilots strategically compensate for display enlargements in surveillance and flight control tasks. *Human Factors*, *48,* 166–181.

Steelman-Allen, K., McCarley, J., Wickens, C. D., Sebok, A., & Bzostek, J. (2009). *N-SEEV: A computational model of attention and noticing*. Manuscript submitted for publication.

Taleb, N. (2007). *The black swan: The impact of the highly improbable.* New York: Random House.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

U.S. Department of Transportation. (2000). *US Department of Transportation aeronautical information manual.* Washington, DC: Federal Aviation Administration.

Varakin, D. A., Levin, D. T., & Fidler, R. (2004). Unseen and unaware: Implications of recent research on failures of visual awareness for human–computer interface design. *Human–Computer Interaction*, *19*, 389–422.

[P] Weintraub, D. J., Haines, R. F., & Randle, R. (1985). Head-up display (HUD) utility: II. Runway to HUD transitions monitoring eye focus and decision times. In *Proceedings of the 29th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 615–619). Santa Monica, CA: Human Factors and Ergonomics Society.

Wickens, C. D., & Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. *International Journal of Aviation Psychology*, *19*, 49–72.

[Q] Wickens, C. D., Alexander, A. L., Thomas, L. C., Horrey, W. J., Nunes, A., Hardy, T. J., & Zheng, X. S. (2004). *Traffic and flight guidance depiction on a synthetic vision system display: The effects of clutter on performance and visual attention allocation* (Tech. Rep. AHFD-04-10/NASA(HPM)-04-1). Savoy: University of Illinois, Aviation Human Factors Division.

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W., & Talleur, D. A. (2003). Attentional models of multi-task pilot performance using advanced display technology. *Human Factors*, *45*, 360–380.

[R] Wickens, C. D., Helleberg, J., & Xu, X. (2002). Pilot maneuver choice and workload in free flight. *Human Factors*, *44*, 171–188.

Wickens, C. D., & Hollands, J. (2000). *Engineering psychology and human performance* (3rd ed). Upper Saddle River, NJ: Prentice Hall.

Wickens, C. D., McCarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., & Zheng, S. (2008). Attention-situation awareness (A-SA) model of pilot error. In D. C. Foyle & B. L. Hooey (Eds.), *Human performance modeling in aviation* (pp. 213–242). Boca Raton, FL: Taylor and Francis/CRC.

Wickens, C. D., Sebok, A., Bzostek, J., Steelman-Allen, K., McCarley, J., & Sarter, N. (2009). NT-SEEV: A model of attention capture and noticing on the flight deck. In P. M. Vidulich and P. Tsang, *Proceedings of the 15th International Symposium on Aviation Psychology* [CD-ROM]. Dayton, OH: Wright State University.

Wickens, C. D., Thomas, L. C., & Young, R. (2000). Frames of reference for display of battlefield terrain and enemy information: Task-display dependencies and viewpoint interaction use. *Human Factors*, *42*, 660–675.

Wiener, E. (1977). Controlled flight into terrain accidents. *Human Factors*, *19*, 171–180.

Christopher D. Wickens is a senior scientist at Alion Science Corporation, Micro Analysis and Design Operations, in Boulder, Colorado, and professor emeritus at the University of Illinois at Urbana-Champaign. He received his PhD in psychology from the University of Michigan in 1974.

Becky L. Hooey is a senior research associate for the San Jose State University Research Foundation at NASA Ames Research Center in Moffett Field, California. She received her master's degree in psychology from University of Calgary in 1995.

Brian F. Gore is a senior research associate at the San Jose State University Research Foundation at the National Aeronautics and Space Administration (NASA) Ames Research Center in Moffett Field, California. He received his master's degree in human factors from San Jose State University in 1999.

Angelia Sebok is a lead human factors engineer at Alion Science and Technology in Boulder, Colorado. She earned her MS in industrial and systems engineering from Virginia Tech in 1991.

Corey S. Koenecke is a senior programmer at Alion Science Corporation, Micro Analysis and Design Operations, in Boulder, Colorado. He received his BS in computer information systems from Metropolitan State College of Denver in 2001.

Date received: March 23, 2009
Date accepted: August 18, 2009