Predicting Pilot Performance in Off-Nominal Conditions: a Meta-Analysis and Model Validation

Pilot response to off-nominal (very rare) events represents a critical component to understanding the safety of next generation airspace technology and procedures. We describe a meta-analysis designed to integrate the existing data regarding pilot accuracy of detecting rare, unexpected events such as runway incursions in realistic flight simulations. Thirty-five studies were identified and pilot responses were categorized by expectancy, event location, and whether the pilot was flying with a highway-in-the-sky display. All three dichotomies produced large, significant effects on event miss rate. A model of human attention and noticing, N-SEEV, was then used to predict event noticing performance as a function of event salience and expectancy, and retinal eccentricity. Eccentricity is predicted from steady state scanning by the SEEV model of attention allocation. The model was used to predict miss rates for the expectancy, location and highway-in-the-sky (HITS) effects identified in the meta-analysis. The correlation between model-predicted results and data from the meta-analysis was 0.72.

Christopher D. Wickens[1], Becky L. Hooey[2], Brian F. Gore[2], Angelia Sebok[1], Corey Koenecke[1], and Ellen Salud[2]

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 entitled NextGen (JPDO, 2007). This program includes defining a set of new technologies and procedures, integrating the flight deck with air traffic management, as supported by various automation tools and decision aids. While the increased airspace productivity fostered by these technologies and procedures has been well modeled and predicted, the safety implications when **unexpected and unpredicted** circumstances prevail are less understood (Burian, 2008). 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 (Gore, 2008; Foyle & Hooey, 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 the one hand, ample data exist to show that people's response to the unexpected slows in inverse proportion to 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. While 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*, (Simons & Levin, 1997; Rensink, 2002; Stelzer & Wickens, 2006), or inattentional blindness. In a classic study of situation awareness breakdowns in aviation, Jones and Endsley (1996) observed that the majority of such breakdowns occurred at the first phase of Situation Awareness (SA) (noticing and perception), rather than 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 TCAS-position broadcast (Command of Aeronautics, 2006) or the unintentional decoupling of an autopilot (Wiener, 1977).

The modeling of pilot response delay (or non-response) to unexpected events is important for projections of NextGen procedural safety because of the time and money required to carry out pilot-in-the-loop (PIL) simulations. Also, manipulations that can be made in PIL simulations may be limited, particularly for conceptual systems and procedures for which pilots may not have experience, and 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 MIDAS; Gore, 2008) that can make predictions about performance in operationally meaningful units (e.g., seconds delayed, events missed) can fill this gap. While such models may not be able to offer precise predictions of optimal configurations, they often can identify poor designs, and can be used to narrow the parameter space that should be examined more thoroughly with PIL research.

Our approach to this issue consists of four phases: 1) Identifying, through a meta-analysis, pilot response parameters (noticing time and miss rate) for unusual events. 2) Developing and refining a computational model (Noticing – Salience, Expectancy, Effort, and Value, N-SEEV) to predict noticing parameters for unexpected events. 3) Validating our model predictions against the meta-analysis data. 4) Upon validating the model, applying it to a series of NextGen scenarios.

[1] Alion Science and Technology: MA&D Operations, Boulder, Colorado
[2] San Jose State University / NASA Ames Research Center, Moffett Field, CA

METHOD: META-ANALYSIS

The aviation human factors literature was thoroughly reviewed from the following sources: Proceedings of the Annual Conference on Manual Control, Proceedings of the Digital Avionics Systems Conference, IEEE Transactions on Systems, Man, Cybernetics: Part A: Systems and Humans and Part C: Applications and Reviews, International Journal of Aviation Psychology, International Symposium on Aviation Psychology, Journal of the Human Factors and Ergonomics Society, 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 four criteria: (1) carried out in a reasonable or high fidelity aviation simulation environment; (2) presented some off-nominal (ON) event or unexpected event, such as an engine failure or runway incursion; (3) were sufficiently descriptive of this event; and (4) presented performance data on the mean time to detect, or the proportion of times it was noticed.

A total of 35 studies that met the above criteria were identified. Within these we made an effort to uniquely categorize each study (or condition within a study), for which off-nominal performance was measured, in terms of five dichotomous variables: (1) whether the ON event was visible out-the-window or head down in the cockpit; (2) whether a Head-Up Display (HUD) was present or not; (3) whether a highway-in-the-sky (HITS) was present or not (Wickens & Alexander, 2009); (4) whether the event was truly surprising (e.g., occurred on the last landing of the experiment within an otherwise failure-free series of landings) or simply "unexpected" (e.g., a failure of one system late in an experiment, but following an earlier failure of a different system); (5) whether the event occurred during taxi, or during high workload periods of flight (particularly approach and landing). The difference described in point 4, the contrast between truly surprising events and simply unexpected events, was referred to as the difference between "black swans" and "grey swans" by Taleb (2007).

RESULTS: META-ANALYSIS

We examined the 32 cells formed by this 2x2x2x2x2 "design" and found that several of them were unpopulated by any valid experimental data; or had too few observations to contribute to reliable estimates of mean response time (RT) or event detection rate. Hence considerable pooling of data across these dimensions was required (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 reliability) a procedure often performed in meta-analyses. This weighting was accomplished by summing the two terms of the ratio (#events detected)/(total #events experienced) across all studies within a cell of the relevant comparison. The pooling procedures eventually yielded three dichotomous contrasts that produced highly reliable statistical effects on detection performance, which was expressed as miss rate (MR):

- (1) Out-the-window (OTW) events versus heads-down location of ON events (MR = 0.55 and 0.19 respectively; $\chi^2 = 21.65$, $p < 0.01$). Outside world events were missed more than events that were presented on a heads-down display.
- (2) Detection of truly surprising OTW ON events while flying with or without a HITS; (MR = 0.55 and 0.26 respectively; $\chi^2 = 20.46$, $p < 0.01$). Truly surprising ON events, present in the outside world, were more likely to be missed when flying with a HITS than when flying without a HITS.
- (3) Detection of OTW events that were truly surprising versus simply unexpected. (MR = 0.50 and 0.23 respectively; ($\chi^2 = 40.79$, $p < 0.01$). Outside events that were truly surprising were less likely to be noticed than simply unexpected events. There was inadequate response time data in most studies, so this analysis focused exclusively on miss rate.

It is noteworthy that these robust effects were observed in spite of the fact that the studies pooled within each category differed on other variables in a manner to increase within-category variance (and hence reduce statistical power). Also, in selecting these particular dichotomies, care was taken to ensure that there were no major confounds between levels (e.g., HITS studies were carried out with novices, non-HITS studies with experts). The variables identified through this phase of the research were then mapped onto and used to populate the SEEV parameters in the N-SEEV portion of the research.

METHOD: N-SEEV MODEL IMPLEMENTATION

The SEEV model (Wickens et al., 2003; Wickens et al., 2008) predicts how visual attention is guided in large scale environments by the **salience** of events, inhibited by the **effort** required to move attention, 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). 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 this model drives the eyeball around an environment, such as the dynamic cockpit, according to the

four SEEV parameters. 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, like maintaining stable flight (Wickens et al., 2008).

The N-SEEV model (Noticing-SEEV; Wickens et al., 2009; McCarley et al., 2009) is an elaboration of SEEV, which 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 value, derived from a computational model of Itti and Koch (2000), and augmented to include the salience of changes. 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, while valuable to be noticed, may be neither expected nor salient. N-SEEV associates these parameters with numeric values, 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, the model will actually predict a *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 additional assumptions that, if the area of the TBNE is not fixated within some criterion time (Tc), then the event will be missed. In this way, the model if run repeatedly, can generate a miss probability estimate.

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 using a Boeing-747 simulator (Sarter et al., 2007) was used to validate the parameters of SEEV. Through this effort, it was possible to predict over 90% of the variance of pilot scanning in the GA studies, and 75% of the variance in scanning within the automated Boeing 747 cockpit (See McCarley et al., 2009; Wickens et al., 2009). Second, the noticing (N) component of the N-SEEV model was validated against noticing time and miss-rate data collected by Nikolic, Orr and Sarter (2004), in an experiment simulating pilots noticing 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 miss rate data from the various conditions of the Nikolic et al. experiment (Wickens et al., 2009). In particular, the assumption of a scan rate of 2 fixations/sec, and a miss criterion of 7.5 s (e.g., 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 obtained data of greater than 0.97 for both noticing time and miss rate.

RESULTS: N-SEEV MODEL VALIDATION

The next step in this effort was to validate the model against the meta-analysis data, 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, the six different scenarios, created by the two levels of each of the dichotomous variables revealed from the meta-analysis (OTW vs. down location, presence or absence of HITS, and high vs. low expectancy) were each characterized by parameters of N-SEEV. Six model runs were then carried out, one for each level of the three dichotomies, with each run iterated 1000 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; Gore et al., 2009), a set of miss rate predictions were generated across the six conditions.

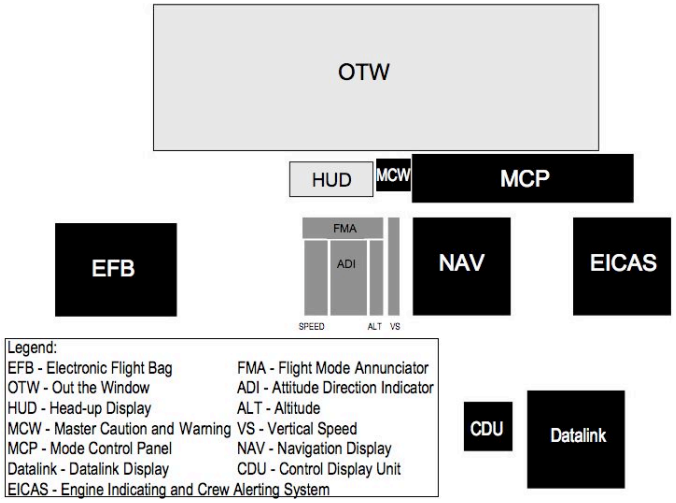

Figure 1. Instrument panel layout upon which model was exercised. The boxes represent cockpit display areas of interest.

For the **HITS comparison**, the ON was located just above the OTW view in Figure 1, and was given a bandwidth (BW) of 0 (no expectancy). The attitude direction indicator (ADI) was assigned high bandwidth and value levels when the HITS was present, and lower levels when the HITS was absent. For the **ON location comparison,** the ON was located either above the OTW view or "down" at the location just below the ADI. Standard (non HITS) BW and value levels were chosen appropriate for visual flight rules (VFR) flight (Wickens et al., 2003) and ON expectancy (BW) while low, was not at 0. For the **expectancy comparison** ON was located above the OTW view, and was assigned a BW value of either 0 (no expectancy) or 0.2 (low expectancy), using other parameters to characterize visual meteorological conditions (VMC) flight.

The model-predicted miss rates for the six conditions are shown on the X-axis of Figure 2. The Y-axis depicts the corresponding obtained miss-rates from the meta-analysis. Connected pairs of points represent the two conditions compared in each of the three contrasts, as labeled (i.e., expectancy effect, ON location effect, and HITS presence effect).

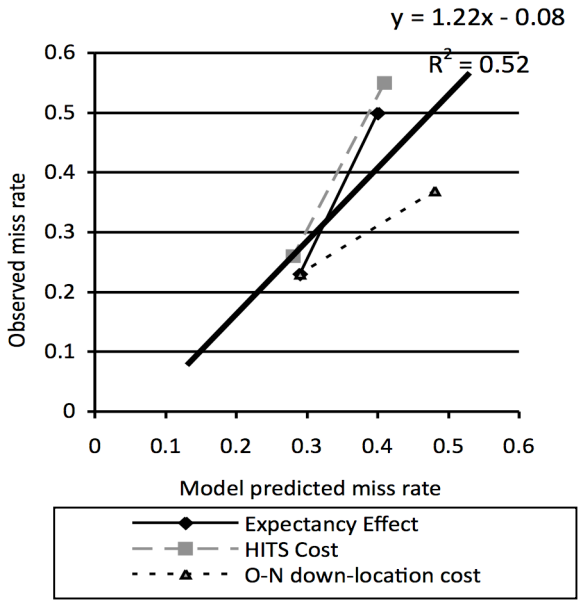

Figure 2. Validation of N-SEEV model predicted miss rate (x-axis) against measured ON miss rate from the meta-analysis. Lines connect the two points within each dichotomous comparison.

Two general features of this model validation are noteworthy. First, the overall correlation, across the six data points, between predicted and obtained miss rate was r = 0.72, a reasonably good 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, a regression line for the six points generate a slope value reasonably close to 1.0 (1.2) and an intercept reasonably close to zero (0.08). This means that not only are changes in model predictions echoed in changes in obtained data (the positive correlation), but the actual value of predicted miss rate corresponds closely to the actual value obtained. Indeed we note that all of the six empirical data points were predicted within 15% (on an absolute scale; that is, for example 55% observed, 40% predicted). Furthermore, four of the data points were predicted within 7%.

PREDICTION OF NEXTGEN PERFORMANCE

The validation reported above provided convincing evidence that N-SEEV could predict existing off-nominal event detection data. From this validation, the next step was to make predictions regarding off-nominal detections associated with different aspects of NextGen technology. We present below a sampling of some of the predicted test model runs and their results to examine the implications of different proposed NextGen procedures on miss rate (MR) for truly surprising ("black swan"), non-salient events at the locations indicated:

- Uplinking taxi clearances during approach, using an electronic flight bag to check taxi routes: Event out the window: MR = 0.51, Event on datalink display: MR = 0.57.

Self-separation responsibility using a cockpit display of traffic information (CDTI) located on the Navigation Display (ND). Event out the window: MR = 0.62, OW event coupled with engine failure management (simulated through cognitive tunneling): MR = 0.83.

- Very closely spaced parallel approaches (VCSPA), and monitoring specialized VCSPA display on the ND. Event on ND: MR = 0.29, Event on EICAS; MR = 0.58.

Miss rate predictions for these and other scenarios can be found in Gore et al. (2009).

SUMMARY AND CONCLUSIONS

In summary, this research has shown how a computational model of noticing can predict data from highly realistic flight simulations, regarding the probability of noticing off-nominal events that are either totally surprising ("black swans") or simply unexpected ("grey swans"). The model predicted the magnitude of three effects of considerable importance to aviation safety: the reduced detection of unexpected events, the "attentional tunneling" effects of the HITS head down on the instrument panel, and the general advantage of detecting rare events in the forward view, rather than head down.

To the extent that the model validation shown in Figure 2 can be trusted, then the extrapolated prediction of the implications of NextGen procedures on miss rate may also be considered as somewhat valid, and can identify areas of potential concern, to be explored by further PIL simulation. The model can be adapted to mimic mitigating strategies, such as a changing display layout, highlighting events, or reducing bandwidth (through automation) of otherwise attention-demanding channels. Such changes can be predicted to reduce the high visual and cognitive workload associated with these procedures; and the model can predict the degree of benefit or improvement in flight deck safety. This research indicates that, even in NextGen airspace, the probability of missing "black swan" events is high (MRs > 0.30). Further work is needed to develop training, procedures, and display designs to support pilots in identifying and responding to these "black swan" events.

We acknowledge an important limitation of the current approach, that arises from the somewhat unusual technique of using the meta-analysis data to validate the model in a "postdictive" fashion. In particular, our approach could be criticized because of the selective set of conditions chosen for validation. Why, for example, did we only examine the HITS effect for black swans, and the location effect for grey swans? We selected these particular levels for comparison because they were well populated with data, and therefore provided strong, statistically reliable effects; and such effects provide a better challenge for model validation, than comparisons in which no differences were found.

Future efforts can build upon the approach presented herein by undertaking human-in-the-loop research to identify where information should be presented to predict noticing times to off-nominal events in NextGen scenarios thereby turning the black swans of the present research into the grey swans of future research.

Acknowledgments

This research was supported by grant # NASA ROA2006: Airspace Systems – NGATS ATM: Airspace Project, NNX08AE87A to San Jose State University (Project PI: Brian.F.Gore@nasa.gov). The authors would like to thank NASA's Technical Monitor (Dr. David Foyle) for his overview of the grant and all reviewers for their comments on the present document.

References

Command of Aeronautics (2006). Final report A-0-22 CENIPA, 2008, Brazil.

Burian, B. K. (2008). Perturbing the system: Emergency and off-nominal situations under NextGen. *International Journal of Applied Aviation Studies*, *8*(1), 114-127.

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 Press.

Gore, B. F., Hooey, B. L., Wickens, C. D., Sebok, A., Hutchins, S., Salud, E., Small, R., Koenecke, C., & Bzostek, J. (2009). Identification of NextGen air traffic control and pilot performance parameters for human performance model development in the transitional airspace. NASA Final Report and Deliverable. ROA 2006: Airspace Systems – NGATS ATM: Airspace Project, NRA # NNX08AE87A, San Jose, CA: San Jose State University.

Gore, B. F. (2008). Chapter 32: Human performance: Evaluating the cognitive aspects. In V. Duffy (ed.), *Handbook of Digital Human Modeling*, Boca Raton, FL: Taylor and Francis / CRC Press.

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.

McCarley, J., Wickens, C. D., Steelman, 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.*

Nikolic, M. I., Orr, J. M., & Sarter, N. B. (2004). Why Pilots Miss the Green Box: How Display Context Undermines Attention Capture. *The International Journal of Aviation Psychology*, *14*(1), 39–52.

JPDO. (2007). Joint Program Development Office. Concept of Operations for the Next Generation Air Transport System. V2: Washington, DC: FAA.

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

Mack, A. & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.

McCarley, J., Wickens, C. D., Steelman, 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.*

Rensink, R. A. (2002). Change detection. *Annual Rev of Psych, 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*(3), 347-357.

Simons, D. J. & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Science, 1*(7), 261-267.

Stelzer, E. M. & Wickens, C. D. (2006). Pilots strategically compensate for display enlargements in surveillance and flight control tasks. *Human Factors, 48*(1), 166-181.

Taleb, N. (2007). *The Black Swan. The Impact of the Highly Improbable.* New York, NY: Random House.

Wickens, C. D. & Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. *International Journal of Aviation Psychology, 19*(2), 1-17.

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*(3), 360-380.

Wickens, C. D., Helleberg, J., & Xu, X. (2002). Pilot maneuver choice and workload in free flight. *Human Factors, 44*(2), 171-188.

Wickens, C. D. & Hollands J. (2000) Engineering Psychology & Human Performance, 3rd Ed. Upper Saddle River, NJ: Prentice Hall.

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

Wickens, C.D., McCarley, J.S., Steelman-Allen, K., Sebok, A., Bzostek, J.. & Sarter, N. (2009). NT-SEEV: A model of attention capture and noticing on the Flight Deck. To appear in the *Human Factors and Ergonomics Society Annual Meeting Proceedings*, Santa Monica, CA: HFES.

Wickens, C. D., Sebok, A., Kamienski, J., & Bagnall, T. (2007). Modeling situation awareness supported by advanced flight deck displays. *Human Factors and Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: HFES.

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