Wickens, C.D., Sebok, A., Gore, B.F., & Hooey, B.L. (2012). Predicting pilot error in NextGen: Pilot performance modeling and validation efforts. Proceedings of the 4th International Conference on Applied Human Factors and Ergonomics (AHFE), July 2012.

# Predicting Pilot Error in NextGen: Pilot Performance Modeling and Validation Efforts

*Christopher D.Wickens[1], Angelia Sebok[1], Brian Gore[2], and Becky Hooey[2]*

*[1]Alion Science and Technology, Boulder, Colorado*

*[2]San Jose State University at NASA Ames Research Center, Moffett Field, California*

*cwickens@alionscience.com, asebok@alionscience.com*
*brian.f.gore@nasa.gov, becky.l.hooey@nasa.gov*

## ABSTRACT

We review 25 articles presenting 5 general classes of computational models to predict pilot error. This more targeted review is placed within the context of the broader review of computational models of pilot cognition and performance, including such aspects as models of situation awareness or pilot-automation interaction. Particular emphasis is placed on the degree of validation of such models against empirical pilot data, and the relevance of the modeling and validation efforts to NextGen technology and procedures.

**Keywords**: Flight deck, pilot models, NextGen, pilot error, error models

## 1 INTRODUCTION

The next generation of air transportation systems will impose a large set of new procedures and new technology on the triad linking the flight deck pilots, the air traffic controllers, and automation. Will such procedures be feasible with the technology provided? Added procedures and responsibility (particularly on the

flight deck) may lead to pilot workload overload. Layers of automation designed to mitigate this overload may degrade situation awareness in a manner that may be of little consequence until an off-nominal event or an automation failure occurs (Burian, 2007; Wickens et al., 2009). In the conventional approach to air space evolution, human-in-the-loop (HITL) simulation is coupled with application of principles of good human factors design to achieve effective development. However, as is often stated, application of good human factors principles rarely guarantees fully-effective design with complex systems, and HITL simulation is often time consuming. Even a well-designed study may cover only a small range of the parameter space that should be investigated, and may involve only 8-10 pilots. While such a sample is often adequate to evaluate routine performance, it does not provide adequate statistical power to infer the nature of pilot response to off-nominal events (Wickens, 2001, 2009); and yet these are the critical break points of aviation safety (Wickens, 2000).

The complementary alternative, which is the focus of the current paper, is the computational model of pilot performance (Foyle and Hooey, 2008). This may be either an analytical equation, or a discrete event simulation that represents some aspect of pilot performance. It may range from the relatively simple, such as a Fitts' Law analytical model predicting pilot reach time to controls, or a model predicting pilot subjective ratings of flight dynamics, to the very complex, which might simulate all aspects of a pilot's vision, cognition and action (e.g., Man-machine Integration Design and Analysis System MIDAS; Gore, 2010).

The objectives of this paper are to describe a project in which we defined criteria for the relevance and usefulness of pilot models to certificating agencies like the FAA, gathered all available literature on pilot models, and then performed an in-depth analysis on models addressing one specific facet of aviation: pilot errors.

## 2 MODEL FEATURES AND CRITERIA

We identified a preliminary set of coding criteria or features to characterize models, based heavily on guidelines set out for models of display and control configuration by Wickens, Vincow, Schopper, and Lincoln (1997). Model features used in this evaluation belonged to one of two general classes. First, there are descriptive features that have no evaluative (e.g., "better" or "worse") connotation to them, such as the type of model (e.g., simulation versus equation). This defines class A. Second, there is the class of features we refer to as "criteria", whose rating for any particular model ***does implicitly or explicitly suggest greater or lesser desirability***. For example, all other factors being equal, a validated model is more desirable than an unvalidated one. This second general class of evaluative features can be further subdivided. One set of evaluative features (class B) are those that are clearly related to the quality of the model validation efforts, including such items as how close the population in the validation experiments is to commercial pilots, as well as the degree of success in the validation effort, in terms of the model's ability to accurately predict pilot performance. The other set of evaluative features (class

C) are those features that are important in assessing the overall value or utility of the model, but are not related to validation. These include features such as usability and software support that were beyond the scope of this project. These original criteria, along with other model features were iteratively refined with team member input, based on lessons learned while reviewing the literature, and specific requirements related to NextGen.

Altogether, nine features were defined to characterize each separate research paper that examined a pilot performance model, with each categorical feature having a number of different levels. We define first the three class A features and then the six class B features.

(A1): *Type of modeling effort*. The modeling papers in the analysis were classified according to model type. The model types included: discrete event simulation, analytic equations, regression, or qualitatively descriptive models.

(A2): *Aspect of pilot performance modeled*. After reviewing the available models, this categorization appeared to encompass all human performance variables evaluated by the model, and each term provided a useful aviation-relevant description that could also be associated with keyword searches. It was important that these categories be not mutually exclusive. For example, a model designed to assess situation awareness could be tailored, in a particular application, to predict errors in situation awareness, and hence might also receive classification as an error model (for that application).

(A3): *Model name*. Several of the models we review depend upon the same fundamental architecture, and are often associated with a particular name, such as ACT-R or MIDAS. Where relevant and available, this is called out as a separate feature of model description, to aid searching and classification.

(B4): *Empirical data available*. Here we determined if empirical human performance data were reported in the paper that could be directly employed (or was employed) to evaluate and validate quantitative predictions of the model.

(B5): *Validation approach*. Validation approach may range from quantitative validation (i.e., product-moment correlation between model prediction and observed data across a range of evaluated conditions) to qualitative validation (e.g., the pattern of errors predicted by the model was quite similar to that shown by the pilots).

(B6): *Correlation result.* These numerical terms describe the results of a quantitative validation approach through the value of the correlation and the sample size. The sample size does not refer to the number of pilots sampled, but rather the number of conditions across which the bivariate point of a [model prediction - data point] could be determined (e.g., three conditions of weather; n = 3).

(B7): *Correlation method*. The most desirable model validation is one that predicts mean pilot performance (e.g., errors, workload) across two or more different conditions, and such difference is captured by the model. A less preferred correlational method correlates performance data to different pilots (not different conditions).

(B8) and (B9): *Population and Test-bed.* These can each be considered evaluative, in the sense that some levels on each of the two features (e.g., population

of commercial pilots, test bed on a high- fidelity commercial aircraft simulator) make the study more realistic and therefore of higher utility than studies employing less representative populations (e.g., general aviation pilots) or test-beds (e.g., PC simulations).

## 3 MODEL IDENTIFICATION AND CLASSIFICATION

Our search identified 160 relevant articles, which were then reviewed and coded according to the Class A and B features listed above. These categories represent a set of factors by which models and validation efforts can be characterized. In some cases, two independent sets of codes were applied to a single article if that article reported two validation experiments of a single model or if the paper presented two models that were validated using different aspects of a single data set.

A sample of 31 coded articles was selected for a round of independent coding to assess inter-rater reliability. Of these 31 articles, there was perfect agreement between the two raters on 28 of the models, indicating a reliability of 90.3%. The three discrepancies were resolved.

## 4 OVERALL MODEL VALIDATION STATISTICS

The number of models that could be fit into each category of pilot performance (e.g., manual control, error) was tallied. A given model validation was sorted into more than one category of pilot performance if it cut across multiple categories (e.g., a model of pilot visual scanning of a flight management system could be coded as both Automation and Vision). The following lists the 13 model aspects and for each, the number of articles identified and the percentage of those articles containing validation information of some sort:: Pilot-automation interaction (12, 25%); Communications (7, 0%); Decision making (22, 27%); Error (25, 28%); Fatigue (3, 67%); Manual control (30, 67%), Multi-task and task management (8, 50%); Procedures (24, 21%); Full-pilot model (i.e., including more than 3 of the aspects) (51, 24%); Situation awareness (12, 58%); Spatial disorientation (1, 0%); Visual/attentional processes (49, 63%); Workload (37, 43%). Full details are provided in Wickens et al. (2011).

The overall statistics of these are that, of 281 different modeling efforts (some of the 160 articles contained more than one effort), 118, or 42%, offered some form of validation.

## 5 PILOT ERROR MODELS

As described above, 25 of the articles focused on pilot error. We now focus on these in greater detail. The choice of pilot *error* as a target for our first in-depth analysis of the results was because of its direct linkage to "show stopping" mishaps and accident (Wiegmann and Shappell, 2003) and because pilot error, however

defined, is an important component in overall system reliability. The 25 computational pilot error models were distilled to 17 separate modeling exercises and were classified into five major subgroups. In the next section, we describe these modeling approaches in greater detail.

5.1 Human Reliability Analysis.

A set of three papers (Salmon et al., 2002; Stanton et al., 2003; and Salmon et al., 2003) evaluated different means of classifying pilot error, with a focus on the Systematic Human Error Reduction and Prediction Approach (SHERPA). The validation is was accomplished by comparing the kinds of errors that SHERPA (and two other comparable error taxonomies) would predict as SHERPA was exercised by a sample of students, with errors that were predicted to occur by expert pilot subject matter experts (SMEs). Both predictions were made within an auto-land flight scenario. A validity score was reported in terms of the hits (errors that were predicted by SHERPA which were also predicted by SMEs) and misses (errors predicted by SMEs not predicted by SHERPA). Scores indicated about 75% validity. That is, 75% of the SME-identified errors were predicted by SHERPA.

A paper by Miller (2001) presented the Aviation Safety Human Reliability Analysis Method (ASHRAM), also heavily founded on the principles of human reliability analysis. The paper presents ways for predicting plausible error inducing conditions on the basis of a three-stage model of pilot information processing (perception, cognition, action), and the influence of performance shaping functions. It can be used prospectively, to predict these error-likely conditions, or retrospectively in mishap analysis, to understand how breakdowns in pilot information processing could have played a role.

## 5.2 Procedural Risk Models

A set of four papers (Stroeve, Blom, and Bakker, 2011; Stroeve, Blom and Bakker, 2009; Stroeve and Blom, 2005; Blom, Corker, Stroeve, and van der Park, 2003) was centered around one generic model – TOPAZ (Traffic Organization and Perturbation Analyzer;), which was developed at NLR (Netherlands Aerospace Research Lab). In these papers, TOPAZ was applied to the prediction of runway incursions resulting from one aircraft failing to stop at an intersection where another was on a take-off run. The focus of the model was to predict objective risks. Thus this model incorporated component models of two pilots and a ground controller in their interaction. Because it was a Monte Carlo simulation model, repeated runs predicted the relative frequency of these very rare events (runway collisions). However the studies report verification, rather than validation. That is, the model allowed users to exercise different environmental, pilot and equipment conditions (e.g., high vs. low surface visibility, presence or absence of alerting systems, different kinds of pilot errors) to examine the influence of these factors on incursion likelihood. Importantly, the most recent paper by Stroeve, Blom, and Bakker (2011) explicitly incorporates automation effects, so that the influence of human automation interaction (HAI)-related factors can be predicted, an element quite clearly related to NextGen.

## 5.3 Knowledge-Based Procedural Models

A set of three pilot model papers, focusing on the sources of failure in retrieving procedural and declarative knowledge from memory were directly based on ACT-R or ACT-R assumptions (see Lebiere et al., 2008, for a description). In ACT-R, the memory for and activation of goals to trigger actions is fallible (e.g., forgetting to activate an automation function). Hence errors are generally memory failures caused by insufficient strength of the goal to generate the action at a particular time, or by a context that activates an inappropriate goal above the threshold where its actions are triggered.

The paper by Fotta et al. (2007) describes an application of the ACT-R based model Human Error Modeling for Error Tolerant Systems (HEMETS) to the design of a fighter-cockpit interface in predicting different forms of errors (e.g., attention, planning, motor).

The paper by Byrne et al. (2008) is one of a set of five papers (four others reviewed below) that modeled taxiway turn errors, based on data provided by NASA Ames. These chapters all appear in the integrative book Human Performance Modeling in Aviation by Foyle and Hooey (2008). In this database, a corpus of twelve errors committed by eighteen two-pilot crews, over three scenarios each in a high-fidelity landing and taxi pilot in the loop (PITL) simulation scenario at Chicago O'Hare Airport was compiled. The error descriptions were provided to the modeling teams, along with extensive other material regarding verbal transcripts before and after touch down, airport surface layout and timing of events (e.g., communications, passage of intersections).

Byrne et al. (2008) focused on decision errors in which a SME generated a set of decision rules that could be applied by pilots to decide when and whether to turn at a particular intersection. These were implemented in ACT-R, along with information about differences in time stress that could lead pilots to use more heuristic rules (faster, but less accurate), or more formal rules (slower to execute, more accurate). When these were incorporated in the ACT-R simulation, errors were generated, and the authors report a qualitative similarity between the pattern of errors generated by ACT-R and those in the error database.

The ACT-R version used by Lebiere et al. (2008, in Foyle and Hooey), applied to the same NASA taxi-turn database, focused to a greater extent on memory errors (factors causing a failure to retrieve the appropriate turn information at each intersection), and hence either turn at the wrong intersection, or fail to turn at the right one.

## 5.4 Error Generation Models

Deutsch and Pew (2008, in Foyle and Hooey) describes a model, the Distributed

Operator Model Architecture (D-OMAR), which employs ACT-R-processes to select procedures (or fail to select procedures). In contrast to the previous two ACT-R versions, which focused on decision and memory errors respectively, this paper explicitly addressed and describes described errors according to the Reason (1990) error taxonomy of slips, mistakes, and violations. Their model produced errors driven by expectation based on partial knowledge (e.g., an incorrect turn driven by the modeled pilot's expectation that the taxi clearance would be the shortest route to the gate) and errors driven by habit (e.g., an incorrect turn driven by the pilot's habit to always turn left to his/her gate despite an unusual taxi clearance directing a right turn).

Air MIDAS (Corker et al, 2008 in Foyle and Hooey) is a full pilot model that has received fairly extensive validation in other model categories (e.g., workload, procedures, visual attention; see Gore, 2010). However one particular model effort was focused on the taxi-turn error data set described above. Environment triggers (e.g., turns, signs, ATC calls) elicited the baseline behaviors that were predictive of human performance in current- day operations. This served to identify risk factors that increase the probability of error or that could mitigate the error. Air MIDAS incorporates many assumptions about working memory, pilot activity scheduling, and mental workload, based on empirical research findings.

The A-SA (Attention-Situation Awareness) model (Wickens et al., 2008, in Foyle and Hooey) focuses on modeling visual attention, and the memory-related loss of situation awareness as factors that lead to either an enhanced or degraded sense of position of where the aircraft is on the taxiway surface, an undesirable state leading to turn errors. To the extent that SA degrades, pilots are left to use data-free decision heuristics (of the sort modeled by Byrne et al., 2008, described above) to decide the direction of turn, and hence make incorrect turns.

The Cognitive Architecture for Safety Critical Task Simulation (CASCaS) is a full pilot performance computational model proposed by Lüdtke et al. (2009). However one particular application is represented to predict two kinds of errors: learned carelessness and cognitive lockup. Both predictions appear to be based on an ACT-R learning mechanism, as pilots have repeated experience with using the flight management system (FMS) in programming particular procedures. As manifested in the error predictions reported in the paper, both types of errors are essentially errors of attention (failure to notice incorrect states) that produce inappropriate actions (or action failures) on the FMS.

## 5.5 Error Detection and Recovery Models

While the previous sections have focused on predicting the occurrence of pilot errors, two final papers focus on the post-error processes of error detection and recovery.

Karikawa et al. (2006) presented the Pilot Cognitive Simulation (PCS), which models full pilot cognitive capabilities. The emphasis of this simulation model was on the pilot's mental model of a particular scenario. The model predicts how well pilots will notice errors with and without automation enhancements to the primary

flight display.

Nikolic and Sarter (2003) present a qualitative flow model of recovery from errors, essentially defining two strategies. A backward strategy tries to "undo" the erroneous action. A forward strategy simply ignores the actions that created the error, but tries to recover performance to the ideal currently desired state. Their description of the two strategies offers the qualitative prediction that the forward strategy will be more likely adopted under time pressure. The model was validated against a set of 38 Aviation Safety Reporting System (ASRS) reports in which the error recovery strategy could be evaluated. Of these, 75% were categorized as forward recovery. The authors did not classify the extent to which these (and not the remaining 25%) occurred under greater time pressure. However, it can be inferred that time pressure was present in most cases (there was urgency when recovering from an error in flight), and hence the prevalence of forward reasoning strategies in the ASRS database represents a form of empirical validation.

# 6 CONCLUSIONS REGARDING PILOT ERROR MODELS

Models of pilot *error* represented a relatively small proportion of the overall body of available pilot models. In the review and synthesis described more fully in Wickens et al. (2011), we offered two reasons for what we saw to be the relatively low rate of validation: Only about half (9 of 17) contained data against which the model could be partially validated, and of these, none contained a true quantitative (e.g., correlational) validation.

First, errors in aviation are, fortunately, fairly rare and it is often difficult to induce them in operational settings with enough frequency to obtain a stable performance target for the models to capture (but see Wickens, Hooey, Gore, Sebok and Koenecke, 2009, for a successful attempt to validate models of perceptual errors, and Nikolic and Sarter, 2003, for successful use of ASRS data). When error rate is low, then differences in error rate (e.g., across conditions, or technology) will be less reliable, and it will be harder then to validate how well a model can predict these differences. Naturally, validating model-predicted differences across extremely rare events, like the actual runway incursions examined by the TOPAZ models, becomes nearly impossible.

Second, errors have multiple internal (pilot-information) causes, such as breakdowns in attention, memory, and procedure selection. Thus a single process error model is asked to predict a data base of errors that are typically related to multiple processes, a challenging endeavor to say the least.

Given this state of affairs, a fruitful approach would be to model the breakdown of processes that are pre-cursors to errors, such as attention failures, poor task management, loss of situation awareness and high workload. These have been categorized (see Wickens et al., 2011) and are the targets of our current efforts.

# ACKNOWLEDGMENTS

This research was supported by the Federal Aviation Administration, (DTFAWA-10-X-80005 Annex 1.11, 05-02; 09-AJP61FGI-0002). The authors wish to acknowledge the invaluable contributions of the entire research team: Ron Small, Steve Peters, John Keller, Shaun Hutchins, and Liana Algarin. Tom McCloy of the FAA, David Foyle of NASA Ames Research Center, and Kevin Jordan of SJSU were the technical monitors of this work.

## REFERENCES

Blom, H.A.P., Corker, K.M., Stroeve, S.H. & van der Park, M.N.J. (2003). Study on the integration of Air-MIDAS and TOPAZ. Nationaal Lucht- en Ruimtevaartlaboratorium (The Netherlands Aerospace Research Laboratory) Contractor Report NLR-CR-2003.

Burian, B. (2007). Perturbing the system: Emergency and off-nominal situations under NextGen. *International Journal of Applied Aviation Studies. 8*, 114-127.

Byrne, M.D., Kirlik A., & Fleetwood, M.D. (2008). An ACT-R approach to closing the loop on computational cognitive modeling. In D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: Taylor & Francis / CRC Press.

Corker, K.M., Muraoka, K., Verma, S., Jadhav, A., & Gore, B.F. (2008). Air MIDAS: A Closed-Loop Model Framework. In D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: Taylor & Francis / CRC Press.

Deutsch, S.E. & Pew, R.W. (2008). D-OMAR: an architecture for modeling multi-task behavior. In D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: Taylor & Francis / CRC Press.

Fotta, M.E., Nicholson, S., & Byrne, M.D. (2007). HEMETS – Human Error Modeling for Error Tolerant Systems. Proceedings of the 14th International Symposium on Aviation Psychology, 204-209.

Foyle, D. C & Hooey, B. L. (2008). *Human Performance Modeling in Aviation*. Boca Raton, FL: Taylor & Francis / CRC Press.

Gore, B. F. (2010). Man-machine integration design and analysis system (MIDAS) v5:. In P. C. Cacciabu, M. Hjalmdahl, A. Luedtke & C. Riccioli (Eds.), Human Modelling in Assisted Transportation. Heidelberg: Springer

Karikawa, D., Takahashi, M., Ishibashi, A..,et al. (2006). Human-Machine System Simulation for Supporting the Design and Evaluation of Reliable Aircraft Cockpit Interface. SICE-ICASE International Joint Conference, pp.55-60.

Lebiere, C., Archer, R., Best, B., & Schunk, D. (2008). Modeling pilot performance with an integrated task network and cognitive architecture approach. In D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. CRC press.

Lüdtke, A., Osterloh, J.P., Mioch, T.,et al.. (2009). Cognitive Modelling of Pilot Errors and Error Recovery in Flight Management Tasks. Proceedings of the HESSD.

Miller, D.P. (2001). Development of ASHRAM: A new human-reliability-analysis method for aviation safety. Proceedings of the 2001 International Symposium on Aviation Psychology. Dayton, OH: Wright State University.

Nikolic, M. & Sarter, N. (2003). Towards a model of error management on highly automated glass cockpit aircraft. Proceedings of the International Symposium on Aviation Psychology.

Salmon, P., Stanton, N.A., Young, M.S., Harris, D., Demagalski, J., Marshall, A., Waldman, T., & Dekker S. (2002). Using existing HEI techniques to predict pilot error: A

comparison of SHERPA, HAZOP and HEIST. Proceedings of the HCI Aero 2002 Conference. AAAI. 129-130.

Salmon, P.M., Stanton, N.A., et al.. (2003). Predicting Design Induced Pilot Error: A comparison of SHERPA, Human Error HAZOP, HEIST, and HET, a newly developed aviation specific HEI method. Proceedings of the HCII Conference. 567-571.

Stanton, N.A., Salmon, P et al. (2003). Predicting Pilot Error: Assessing the Performance of SHERPA. Proceedings of the HCII Conference. 587-591.

Stoeve, S., Blom, H., & Bakker, G. (2011) Contrasting safety assessments of a runway incursion scenario by event sequence analysis versus multi-agent dynamic risk modeling. 9th USA/Europe ATM R&D seminar.

Stroeve, S. & Blom, H. (2005). Human performance modeling for accident risk assessment of active runway crossing operation. NLR-TP-2005-428. Technical Report from the Netherlands National Airspace Laboratory.

Stroeve, S., Blom, H. & Bakker G (2009) Systemic accident risk assessment in air traffic by Monte Carlo simulation. Safety Science. 47, 238-249.

Wickens, C.D. (2000). The tradeoff of design for routine and unexpected performance: Implications of situation awareness. In D.J. Garland & M.R. Endsley (Eds.), Situation awareness analysis and measurement. Mahwah, NJ: Lawrence Erlbaum

Wickens, C.D., (2001). The Psychology of Surprise. Keynote address; In R, Jensen (Ed). Proceedings 2001 International Symposium on Aviation Psychology: Columbus Ohio: Ohio State University.

Wickens. C.D. (2009). The psychology of aviation surprise: an 8 year update regarding the noticing of black swans. In J, Flach & P. Tsang (Eds). *Proceedings 2009 Symposium on Aviation Psychology*: Dayton Ohio: Wright State University. (Keynote address).

Wickens, C.D. Hooey, B. Gore, B.F., Sebok, A. & Koenicke, C. (2009) Identifying Black Swans in NextGen: Predicting Human Performance in Off-Nominal Conditions. Human Factors. 51, 638-651.

Wickens, C.D., McCarley, J.S.,et al.. (2008). Attention-Situation awareness (A-SA) model of pilot error. In D. C. Foyle & B. L. Hooey (Eds.) In D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: Taylor & Francis / CRC Press.

Wickens, C.D., Sebok, A., Peters, S., Small, R., Keller, J., Hutchins, S., Algarin, L. Gore, B. F. and Hooey, B. L. (2011). Modeling and Evaluating Pilot Performance and Human Error Potential in NextGen (Phase 1, Interim Report). HCSL Technical Report (HCSL-11-03). Moffett Field, CA: NASA Ames Research Center.

. Wiegmann, D. & Shappell, S,. (2003) A human error approach to aviation accident investigation. Burlington Vt.: Ashgate