Review of Pilot Performance and Pilot-Automation Interaction Models in Support of Nextgen

Angelia Sebok

Christopher D. Wickens

REVIEW OF PILOT PERFORMANCE AND PILOT-AUTOMATION INTERACTION MODELS IN SUPPORT OF NEXTGEN

Angelia Sebok & Christopher D. Wickens
Alion Science and Technology
Boulder, Colorado

Computational models of aircraft pilot performance will gain importance over the next decades, as major evolutions in the national airspace continue to emerge with the NextGen program. Evaluating new technology, or procedures such as self-separation, requires time and resource-consuming pilot-in-the-loop (PITL) simulations. Models can augment PITL findings and they can help to constrain the scope of PITL simulations. If they are validated, such computational models may actually answer some design questions in place of PITL simulations. This paper summarizes a review of modeling efforts to address pilot performance, and elaborates on pilot-automation interaction models.

In the transition to NextGen operations, a major concern is identifying and evaluating potential concepts well before they are put into operation. One approach is through the use of computational modeling to predict operator performance, or human performance modeling. Computational modeling provides a means of predicting performance and evaluating numerous "what if" situations, and is thus particularly useful for evaluating to-be-built systems. In addition, models need to make valid predictions of operator performance. This requires comparing model predictions with actual human performance in real or (for to-be-built systems) simulated conditions.

Methods

To identify the scope of existing pilot performance models and their associated validation efforts, we searched more than 40 potential sources (e.g., the Human Factors Society Proceedings, Human Factors Journal, International Symposium for Aviation Psychology Proceedings, International Journal for Aviation Psychology) to identify papers that described either a modeling effort for predicting pilot performance, or an empirical study to validate such a model. Initially, we identified approximately 500 papers. Upon closer inspection, we were able to eliminate approximately two thirds of these papers as duplicates of other articles, air traffic control (ATC; not pilot) related, or model descriptions without provision of specific model predictions. This left a final set of 187 references. We reviewed these papers to characterize the modeling efforts and compare across the studies. We identified a set of criteria by which to evaluate the models, including descriptive and evaluative features. Descriptive features include the name and type of model (e.g., simulation, analytical), the specific aspect of pilot performance that was modeled (e.g., pilot-automation interaction, communication, error). Evaluative features included 1) whether or not empirical, PITL data were provided to validate the model predictions, 2) whether the empirical data provided correlations (or other quantitative evaluations), or were qualitative in nature, 3) the participants in the study (e.g., professional pilots, college students), 4) the test bed (e.g., flight simulator, desktop flight simulator with mouse and keyboard, or other). The descriptive features allowed us to distinguish the different modeling efforts, and the evaluative criteria provided data to compare the extent to which validation studies had been performed.

After this initial assessment, we conducted five separate deep dive analyses to examine in detail the state of the art of modeling efforts of particular relevance to NextGen operations: Pilot-automation interaction (PAI), error, workload and multitasking (see a companion paper, Wickens & Sebok, 2013, for details), situation awareness, and roles and responsibilities. The deep dive analyses identified how the models predict pilot performance and included a review of the verification and validation efforts for these specific topic areas. As implicitly stated previously, validation was considered to be a comparison of model predications against data gathered in an empirical PITL study. In contrast, verification efforts included subject matter expert (SME) reviews of model predictions for "sensibility" of results, or researchers' own interpretation of the results.

Overview of Model Review Results

We carried out an extensive analysis of the extent to which each modeling effort ***in the deep dive analysis*** had been verified and, if verified, also validated; by assigning ratings to the levels of verification and validation (See

Wickens et al., 2013 for details). From this analysis, we concluded that only 5% of modeling efforts included no discernible attempt at verification or validation. Nearly 40% of efforts included verification efforts, and 31% included qualitative evaluations of predictions against empirical data. Twenty-five percent of modeling efforts included human in the loop simulation studies to provide empirical data for comparison with model predictions. This situation leaves room for improvement, and the analysis identifies clear gaps. Studies frequently focused on a small subset of model predictions, rather than the full range of predictions.

Pilot-Automation Interaction Models

Overview

In NextGen operations, new technologies and capabilities are required to provide a significantly increased volume of operations. One of the key features envisioned to enable integration of these capabilities into the aviation system is a greater reliance on automation. As pilots' tasks expand to include maintaining separation from surrounding aircraft, negotiating trajectories with ATC, and monitoring weather and wake vortex conditions, automation is expected to provide pilots with the support needed to perform these tasks. Thus, models that predict pilot performance when using different types of flight deck automation are highly relevant to NextGen operations. Since "pilot-automation interaction on the flight deck" is a broad area, modeling efforts evaluated specific aspects of the domain. The models reviewed here were diverse, some focusing on specific equipment such as the Flight Management System (FMS), some on particular high-workload phases of flight, and others on specific action sequences used in programming automation.

Architectures Used for Pilot-Automation Interaction

We distinguish a modeling architecture as a software tool and / or a theoretical framework that serve as the basis for specific modeling efforts. A successful validation of one modeling effort provides support for the underlying architecture, but it does not "validate" that architecture. If an architecture addresses a specific aspect of performance (e.g., visual scanning), and one model using the architecture has been shown, through comparison with empirical data to be valid, this finding does lend support for other, similar models developed using that same architecture. Three architectures appeared repeatedly in the review of PAI models, described below.

The Adaptive Control of Thought – Rational (ACT-R) is a unified theory of cognition that integrates theories of attention, cognition, and motor actions (Anderson & Lebiere, 1998). ACT-R models cognition through "production rules" or goal-directed behavior, implemented through a series of "if-then" rules. It includes perceptual inputs and motor outputs. ACT-R's main components are modules, buffers, and a pattern matcher. ACT-R uses perceptual-motor modules (visual and manual modules) to simulate interaction with the physical environment. ACT-R models simulate declarative and procedural knowledge. In addition ACT-R models the actual time required for cognitive steps (e.g., retrieving an item from declarative memory) or implementing an action (e.g., shifting gaze, selecting an item on a display). Thus it readily models procedural activities such as programming an FMS.

The CASCaS (Cognitive Architecture for Safety Critical Task Simulation) architecture, like ACT-R, is a cognitive architecture, which provides a structure and set of rules for simulating human cognition (Lüdtke et al., 2009). CASCaS divides cognitive processes and errors into three different levels, depending on operator experience with a particular task: the autonomous, the associative, and the cognitive level which correspond, respectively, to the skill-based, rule-based, and knowledge-based levels of behavior defined by Rasmussen (1983).

The SEEV and N-SEEV models (Wickens & McCarley, 2008; Steelman-Allen et al., 2011) of visual scanning and noticing predict that attention within a given visual field (e.g., a flight deck) is driven by bottom-up factors of display ***salience*** and ***effort*** (distance between displays), and the top-down factors of ***expectancy*** (bandwidth) and ***value*** or importance of the display for the task. SEEV predicts visual scanning behavior. N-SEEV (Noticing SEEV) uses SEEV to predict scanning and **noticing** of discrete events within the cockpit.

Modeling architectures and efforts that address PAI take a variety of approaches in predicting pilot performance. Models predict performance based on pilot visual scanning and noticing, the time required to complete tasks, workload, and automation induced errors. PAI models have also been applied to design tools and proposed as a basis for adaptive automation systems. These predictions and application areas are described below.

Modeling Approaches to Predicting Performance

Attention / Noticing and Visual Scanning. Several modeling efforts addressed PAI in terms of predicted noticing of important changes on the flight deck (e.g., flight mode annunciator indications on the FMS), or in terms of visual scanning. These efforts all start with the premise that a pilot has to notice an indication to be able to interpret and respond to it, so noticing is a necessary (but not sufficient) first step that a pilot must perform. Boehm-Davis et al. (2002) used an ACT-R model to predict pilot noticing of automation mode changes. The model predicted that pilots were more likely to notice mode changes that were initiated by the pilot, rather than the automation. The authors note that similar trends were observed in previously gathered empirical data.

The SEEV and N-SEEV models of visual attention and noticing have been used to predict pilot noticing on the flight deck (Sebok et al., 2012). The SEEV and N-SEEV models have been empirically validated in previous efforts (Wickens et al., 2008; Steelman-Allen et al., 2011). In each of these efforts, model predictions in aviation flight deck contexts, including a high fidelity Boeing 747-400 simulator (Sarter et al., 2007) were found to predict empirical data of scanning and noticing behavior with correlations above 0.60.

CASCaS was used to predict visual scanning behavior, dwell times in areas of interest on an advanced FMS, and the time required to notice specific visual indications in the cruise and approach phases of flight (Lüdtke et al., 2012). The authors indicated that the *overall* correlation between model predictions and empirical data was high (r^2=0.85). For specific aspects of pilot performance, the model predictions were reasonably accurate. Predicted average dwell times on a display in three phases of flight closely paralleled by empirical data. Similarly, average noticing times for specific visual indications in two phases of flight were predicted to be approximately 1 s in each phase, and were found to be 0.8 s and 1.2 s. These results indicate that the model does a reasonable job of approximating pilot behavior. One concern is that validation data provided are for highly specific tasks or visual areas, yet the operational context includes many tasks and areas.

Polson and Javaux (2001) present a model that predicts why pilots do not often scan the flight mode annunciators, a major issue in FMS monitoring. They apply a Goals, Operators, Methods and Selection rules (GOMS; Card et al., 1983) modeling analysis that, among other features, highlights differences in task priorities in multi-tasking, to predict why this task should be of lower priority when other sources of redundant, equivalent information are available. The authors describe a qualitative evaluation of the similarity between their predictions and the data on FMS monitoring by Huttig, Anders & Tautz (1999).

Time to Complete Tasks. CASCaS was also used to predict time to to handle an uplink from ATC in the cruise and approach phases of flight (Lüdtke et al., 2012). The model predicted that the uplink would require approximately 1 minute, with slightly longer times in the approach phase than in cruise. An empirical study of those conditions revealed that pilots performed the uplink faster during approach than in cruise. No quantitative data were provided. Discussion with pilot SMEs provided insights into the reversal between model predictions and data, explaining that, during the approach phase, pilots typically have to work faster just to get everything done.

Manton and Hughes, 1990, developed a regression equation, based on previously-gathered empirical data, to predict the time to complete tasks using a Multi-Function Keyset (MFK) on the S-70B-2 Seahawk Helicopter, used by the Royal Australian Navy. The MFK, much like an FMS, includes a special purpose keyboard and an 8-line alphanumeric display, used to enter data into or view data contained in a tactical database. The equation predicts time as a function of the number of key presses required, operator pauses, and page changes. Using a stepwise regression, the authors found that the equation predicts 79% of the variance in the data ($p < 0.001$). The authors propose that the model can be used to evaluate different types of automation and system configurations.

Air Man-Machine Integration Design and Analysis System (MIDAS, v1; Pisanich & Corker, 1995) was used to predict which type of FMS automation pilots would use to perform a descent based on the time available to implement the clearance and the modality in which the clearance was delivered (voice or datalink). Three types of automation were considered: an autoload capability (the most highly automated), a CDU, and an MCP (mode control panel, the least automated). The Air MIDAS model predicted that the less time available to implement a clearance, the more likely pilots were to use a less-automated mode. Further, the model predicted that pilots were more likely to select the less-automated modes if a clearance was given by voice than by datalink. While the model was validated against a PITL simulation, the results of that validation could not be easily interpreted because of the use of inappropriate t-test statistics.

Workload. Another approach to predicting pilot performance with automation uses workload. Gil et al., 2009, used enhanced (E)-GOMS to model pilot performance when working with a flight control panel (FCP), a control display unit (CDU) or an enhanced CDU. They predicted workload based on the complexity of the procedures, including the number of submethods being performed, the number of steps needed to complete the submethods, the chunks of information that pilots needed to remember, and the number of information transactions. As complexity increases, so does workload. The authors ran the model for each of the three types of automation and collected data on the complexity indices that varied across automation types. In an empirical study, they gathered four different measures of workload: heart rate, subjective workload (NASA-TLX predictions), vertical flight path deviations, and lateral flight path deviations. They calculated the Spearman correlations for the different complexity indices and empirical performance data, and identified positive and significant ($p \leq 0.05$) correlations between the model predictions and heart rate, and between model predictions and vertical flight path deviations.

Automation-Induced Errors. CogTool (John et al., 2009) is based on ACT-R code, and models the time to complete tasks, errors made on task steps, and failure to complete task steps. In their research, the authors identified three sequential tasks associated with entering a landing speed into the CDU, a critical interface between the pilot and the FMS. They ran their model to predict errors, and iterative improved the model. CogTool accesses a latent semantic analysis (LSA) corpus of terms to predict if pilots will understand the terminology on the CDU. During their first model run, they identified that no pilots would be able to complete the first step of the procedure because they did not understand the terms. The LSA corpus represented a college student's knowledge, not the specialized knowledge that a pilot would possess. By switching to an aviation-specific corpus, the researchers obtained a 10 percent success rate on the first task. A series of other changes were implemented to account for pilots' specialized knowledge, and the model eventually predicted success rates of 92% for the entire procedure. This was considered reasonably accurate, based on one of the author's experience as a pilot who trains new pilots to use the FMS, but it was not validated against PITL simulation data.

Schoppek & Boehm-Davis (2004) used ACT-R to create a model (ACT-Fly) to model pilot awareness, cognition, and errors. They evaluated pilot use of automation at the end of the cruise phase of flight until the initial approach fix. ACT-Fly predicted when pilots would choose a more automated mode (VNAV) or a less-automated mode (FLCH and V/S) in two scenarios. In summary, the model predictions were not well supported by empirical findings. Two scenarios showed 20% and 60% agreement in terms of predicted mode selection. The model did predict the types of errors pilots would make (errors of omission and commission), but the model incorrectly predicted error recovery. Actual pilots were able to recover from their errors. Gil et al. (2009) indicate that their E-GOMS model can predict error, by identifying when the number of chunks to be held in working memory exceeds 5. This approach is based on limitations of working memory (Miller, 1956). No error predictions were made, however.

CASCaS was used by Lüdtke et al., (2009) to predict cognitive errors such as Learned Carelessness, which occurs when pilots routinely perform procedures with multiple steps included to ensure safety criteria are met. If these steps typically do not identify safety concerns, pilots learn that they improve efficiency by skipping these steps. The problem is that sometimes these unsafe conditions do exist, and, by skipping those steps, pilots may sacrifice safety for efficiency. Lüdtke & Osterloh (2010) used CASCaS to predict learned carelessness in a flight re-planning procedure. The researchers modeled a flight condition in which a pilot was repeatedly given ATC clearances that required verification. The model predicted that the pilot would, over time, begin neglecting these checks. An actual pilot performed the same conditions, and – as the model predicted – quit performing the verifications. However, unlike the model, the pilot resumed checking after receiving a supposedly related prompt. The researchers updated their model to include contextual factors (strengthening or inhibiting associations between elements in memory). The updated model then correctly predicted (in 23 of the 24 trials) when the pilot checked the vertical view.

Applications of PAI Models within Design Tools. Three papers described efforts to use PAI models in computerized design tools. These efforts used different types of models, but all had the same goal of helping aviation designers identify and avoid potential design problems. One effort (Gonzales-Calleros et al., 2010) evaluated the FMS interface design for adherence to human factors standards such as font type and color contrast between text and background. The paper outlined an approach to include a cognitive model of pilot performance, but the model was not actually integrated with the evaluation tool.

The Automation Design Advisor Tool (ADAT; Sebok et al., 2012) evaluates and compares potential FMS designs. This effort included multiple analytic models to assess design quality based on human factors principles. The analytic models evaluated design issues of 1) information layout, 2) noticeability of changes, 3) meaningfulness of terms, 4) confusability of terms and symbols, 5) complexity of system design (e.g., modes), and 6) procedures necessary to program the FMS. ADAT included attention models (Wickens & McCarley, 2008), described above, to predict pilot scanning behavior and noticing of FMS mode changes.

A third design tool, CogTool (John et al., 2009), allows a designer to create a "use-case storyboard" with a graphical user interface (GUI), and predict time to complete task or errors made. The GUI is connected with an underlying cognitive model, so the planned sequence of actions on the interface is associated with steps such as noticing and interpreting. These steps are then used to identify the time to complete tasks, and the likelihood of the user selecting the correct action in the sequence.

Summary of Pilot-Automation Interaction Models

In summarizing, we note that we did not include models of adaptive automation, because we found no efforts in which automation was adapted (rather than using automation to adapt an interface). There are many ways to model pilot-automation interaction and predict performance on the flight deck including attention and noticing changes, the design of the automation (interface, interaction), the tasks the pilot performs when using the automation, errors that the pilot can potentially commit. Because there are so many factors that can have an influence, it is difficult to capture all in a single model. The ADAT project integrates several process models applicable to the FMS in software tool. However, to date, the CASCaS effort (Lüdtke et al., 2012) appears to be the most comprehensive type of pilot performance model, addressing attention, interaction, and errors.

Conclusions

Several additional points require mentioning. One of the main findings of this review is that human performance modeling provides a viable tool for predicting pilot performance in to-be-built systems. While models typically focused on limited aspects of performance, we did note that many of the models made predictions that offered insights into potential difficulties with both existing and not-yet-developed systems. Models frequently provided useful data for comparing across conditions, and – even when predictions were incorrect – the models offered insights into pilot cognition and behavior that would have been difficult to learn otherwise. In addition, the vast majority (95%) of modeling efforts included some form of verification or validation. We believe that further efforts should be made to develop standards and guidelines for verification and (particularly) empirical validation, to support the development of more realistic and credible human performance models.

Acknowledgements

The authors would like to thank our colleagues Dave Foyle, Brian Gore, Becky Hooey, John Keller, Ronald Small, Steve Peters, Liana Algarin, and Shaun Hutchins for their contributions to this work. We also thank Tom McCloy and Michelle Yeh from the FAA for their contractual support of this effort. This work was performed under contract DTFAWA-10X-800, 05 Annex 1.11, Task Number 05-02; 09-AJP61FGI-0002.

References

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Boehm-Davis, D.A., Holt, R.W. et al. (2002). Developing and Validating Cockpit Interventions based on Cognitive Modeling. In W.D. Gray & C.D. Schunn (Eds.) *Proceedings Twenty-Fourth Annual Conference of the Cognitive Science*, 27.

Card, S., Moran, T.P. & Newell, A. (1983). *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Associates.

Gil, G., Kaber, D, et al. (2009). Modeling pilot cognitive behavior for predicting performance and workload effects of cockpit automation. *Proceedings 2009 International Symposium on Aviation Psychology*. Dayton, OH: Wright State U., 124-129.

Gonzales-Calleros, J., Vanderdonckt, J. et al., (2010). Towards Model-Based AHMI Development. *EICS '10*. June 21-23, Berlin, Germany.

Hüttig, G., Anders, G., & Tautz, A. (1999). Mode Awareness in a modern Glass Cockpit– Attention Allocation to Mode Information. Paper presented at the *10th Intl. Symposium on Aviation Psychology*, Columbus, OH.

John, B.E., Blackmon, M.H., et al.. (2009). Rapid Theory Prototyping: Example of an Aviation Task. *HFES 53rd Annual Meeting*. *53*(12), 794-798.

Lüdtke, A. & Osterloh, J-P. (2010). Modeling Memory Effects in the Operation of Advanced Flight Management Systems. *Human Computer Interaction Aero Conference 2010*, Cape Canaveral, FL.

Lüdtke, A., Osterloh, J.P., & Frische, F. (2012). Multi-criteria evaluation of aircraft cockpit systems by model-based simulation of pilot performance. *Embedded Real Time Software and Systems Conference*. Feb 1-3, Toulouse, France.

Lüdtke, A., Osterloh, J-P., Mioch, T., Rister, F., Looije, R. (2009). Cognitive Modelling of Pilot Errors and Error Recovery in Flight Management Tasks. *Proceedings of the HESSD*.

Manton, J.G., Hughes, P.K. (1990). Aircrew tasks and cognitive complexity. Paper presented at the *First Aviation Psychology Conference*, Scheveningen, Netherlands.

Miller, G. A. (1956). The magical number seven, plus or minus two. *Psychological Review* 63 (2): 81–97.

Pisanich, G.M. & Corker, K.M. (1995). A Predictive Model of Flight Crew Performance in Automated Air Traffic Control and Flight Management Operations. *International Symposium on Aviation Psychology*.

Polson, P.G., & D. Javaux (2001). A model-based analysis of why pilots do not always look at the FMA. *Proceedings of the 11th Intl Symposium on Aviation Psychology*. Columbus, OH: The Ohio State Univ.

Raeth, P.G., Reising, J.M. (1997). A model of pilot trust and dynamic workload allocation. *Proceedings of the 1997 IEEE National Aerospace and Electronics Conference* (NAECON), July 14-18.

Rasmussen, J. (1983). Skills, rules, knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, 13, 257-266.

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.

Schoppek, W. & Boehm-Davis, D.A. (2004). Opportunities and Challenges of Modeling User Behavior in Complex Real World Tasks. *MMI-Interaktiv*, 7, June. ISSN 1439-7854.

Sebok, A., Wickens, C., Sarter, N.et al. (2012) The Automation Design Advisor Tool (ADAT). *Human Factors and Ergonomics in Manufacturing and Service Industries*. *22*(5), 378-394.

Steelman-Allen, K., McCarley, J. & Wickens, C.D (2011) Modeling the control of attention in visual workspaces. *Human Factors*, 53, 142-153

Wickens, C.D. & McCarley, J.S. (2008). *Applied Attention Theory*. New York: CRC Press, Taylor & Francis Group.

Wickens, C.D., McCarley, J.S., Alexander, A.L., Thomas, L.C., Ambinder, M. & Zheng, S. (2008). Attention-Situation awareness model of pilot error. In D.C. Foyle & B.L. Hooey (Eds.) *Human Performance Modeling in Aviation*. CRC press.

Wickens, C.D. & Sebok, A. (2013). Flight Deck Models of Workload and Multi-Tasking: An Overview of Validation. *Proceedings of the International Symposium on Aviation Psychology*.

Wickens, C.D., Sebok, A., et al. (2013). *Modeling and Evaluating Pilot Performance in NextGen - Final Report*. FAA, Contract DTFAWA-10X-800, 05 Annex 1.11, Task Number 05-02; 09-AJP61FGI-0002.