NASA/TM—2013–216504

Modeling and Evaluating Pilot Performance in NextGen: Review of and Recommendations Regarding Pilot Modeling Efforts, Architectures, and Validation Studies

Christopher Wickens
Angelia Sebok
John Keller
Steve Peters
Ronald Small
Shaun Hutchins
Liana Algarín
Alion Science and Technology, Boulder, CO

Brian F. Gore
Becky L. Hooey
San Jose State University, San Jose, CA

David C. Foyle
NASA Ames Research Center, Moffett Field, CA

April 2013

Since it's founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NASA Aeronautics and Space Database and its public interface, the NASA Technical Report Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include creating custom thesauri, building customized databases, and organizing and publishing research results.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at *http://www.sti.nasa.gov*

- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA STI Help Desk at (301) 621-0134

- Phone the NASA STI Help Desk at (301) 621-0390

- Write to:
  NASA STI Help Desk
  NASA Center for AeroSpace Information
  7121 Standard Drive
  Hanover, MD 21076-1320

NASA/TM—2013–216504

Modeling and Evaluating Pilot Performance in NextGen: Review of and Recommendations Regarding Pilot Modeling Efforts, Architectures, and Validation Studies

Christopher Wickens
Angelia Sebok
John Keller
Steve Peters
Ronald Small
Shaun Hutchins
Liana Algarín
Alion Science and Technology, Boulder, CO

Brian F. Gore
Becky L. Hooey
San Jose State University, San Jose, CA

David C. Foyle
NASA Ames Research Center, Moffett Field, CA

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California

April 2013

Acknowledgements

This research was supported by the Federal Aviation Administration, (DTFAWA-10-X-80005 Annex 1.11). Dr. Tom McCloy is the FAA point of contact, and Dr. David Foyle is the NASA point of contact for this work. Dr. Christopher Wickens is the technical point of contact for this work. The authors would like to thank Dr. Barbara Burian of NASA Ames Research Center for her efforts coordinating the work under the agreement, and Dr. Kevin Jordan of San Jose State University who coordinated the subcontract to Alion Science and Technology.

Available from:

NASA Center for AeroSpace Information
7121 Standard Drive
Hanover, MD 21076-1320
(301) 621-0390

This report is also available in electronic form at: http://www.sti.nasa.gov or http://ntrs.nasa.gov

Table of Contents

Acronyms and Abbreviations

Term	*Meaning*
4D	Four dimensional
A/SA	Attention – Situation Awareness
ACT-R	Atomic Components of Thought – Rational
ADAT	Automation Design Advisor Tool
ADS-B	Automatic Dependent Surveillance-Broadcast
AHMI	Airborne Human Machine Interface
ALARMS	Alerting And Reasoning Management System
ANOVA	Analysis of variance
AOI	Areas of interest
ASHRAM	Aviation Safety Human Reliability Analysis Method
ASM	Actual situation model
ATC	Air Traffic Control
ATM	Air Traffic Management
BC	Between conditions
CAD	Computer aided design
CASCaS	Cognitive architecture for Safety Critical Task Simulation
CDTI	Cockpit Display of Traffic Information
CDU	Control Display Unit
CHESS	Crewstation Human Engineering Software System
ConOps	Concept of operations
CS POC	Civil servant point of contact
CSA	Computational model of situation awareness
CSPA	Closely Spaced Parallel Approach
CTAS	Center TRACON Automation System
CWS	Conditions within subjects
D-OMAR	Distributed Operator Model Architecture
EFB	Electronic Flight Bag
EGOMS	Enhanced Goals, Operators, Methods, Selection
EGOMSL	Enhanced Goals, Operators, Methods, Selection - Language
EVO	Equivalent Visual Operations
EVS	Enhanced Vision System
FAA	Federal Aviation Administration
FCP	Flight Control Panel
FDOF	Flight Deck of the Future
FLCH	Flight level change
FMS	Flight Management System
GA	General aviation
GOMS	Goals, Operators, Methods, Selection
GOMSL	Goals, Operators, Methods, Selection – Language
HAI	Human-automation interaction
HEMETS	Human Error Modeling for Error Tolerant Systems
HIM	Hazard and integrity monitoring
HITL	Human in the loop
HITS	Highway in the sky
HPM	Human performance model

Term	*Meaning*
HUD	Head-up display
HUMAN	Model-based analysis of human errors during aircraft cockpit design
IAN	Integrated alerting and notification
IDM	Integrated decision model
ILS	Instrument landing system
IMPRINT	Improved Performance Research Integration Tool
IPME	Integrated Performance Modeling Environment
ISM	Internal situation model
LOFT	Line oriented flight training
LSA	Latent semantic analysis
MCP	Mode control panel
MFK	Multifunction keyset
MIDAS	Man-machine integration design and analysis system
MRM	Multiple resource model
MRT	Multiple resource theory
NAS	National airspace system
NASA	National Aeronautics and Space Administration
NASA TLX	NASA Task Loading Index
NAV	Navigation
NextGen	Next Generation Air Transportation System
NGOMSL	Natural language goals, operators, methods, selection
NLR	Netherlands Aerospace Research Lab
N-SEEV	Noticing, salience, expectancy, effort, value
OCM	Optimal control model
OPSAMS	Operational procedures safety analysis and monitoring system
OTW	Out the window
PAI	Pilot-automation interaction
PCS	Pilot cognitive simulation
PF	Pilot flying
PFD	Primary flight display
PITL	Pilot in the loop
PM	Pilot monitoring
PNF	Pilot not flying
R&R	Roles and responsibilities
RAAS	Runway awareness and advisory system
RNAV	Area navigation
RNP	Required navigational performance
RPD	Recognition primed decision
SA	Situation awareness
SAC	Subjects and conditions
SAGAT	Situation awareness global assessment technique
SAMPLE	Situation awareness model for pilot-in-the-loop evaluation
SART	Situation awareness reporting technique
SDO	Super density operations
SE	Situation element
SEEV	Salience, expectancy, effort, value
SHERPA	Systematic Human Error Reduction and Prediction Approach
SME	Subject matter expert

Term	*Meaning*
SOAR	State operator and result
SOP	Standard operating procedure
STL	Sample task time-line
SVS	Synthetic vision system
SWARMM	Smart whole air mission model
SWAT	Subjective workload assessment technique
TA	Time available
TCAS	Traffic collision avoidance system
TDM	Total demand model
TLAP	Time-line analysis procedure
TOD	Top of descent
TOPAZ	Traffic Organization and Perturbation Analyzer
TR	Time required
UCM	Undifferentiated capacity model
V/S	Vertical speed
VACCPM	Visual, auditory cognitive spatial, cognitive verbal, psychomotor, speech
VACP	Visual, auditory, cognitive, psychomotor
VCSPA	Very closely spaced parallel approach
VMC	Visual meteorological conditions
VNAV	Vertical navigation
VSD	Vertical situation display
WINDEX	Work load index
WV	Wake vortex

Executive Summary

The effort described in this report was a project to support the Federal Aviation Administration (FAA) in evaluating and comparing modeling approaches to predict pilot performance in NextGen operations. This research effort was intended to assess the current state of the art regarding modeling of pilot performance on the flight deck and to provide guidance regarding research needs in modeling and validation. One hundred and eighty-seven references were identified that examined computational models of pilot performance. We identified 12 different features of each of these modeling efforts to facilitate comparisons. A subset of these features focused on the quality and extent of validating model predictions against pilot-in-the-loop simulation data. In addition, we identified 12 different aspects of pilot performance that were the focus of the model in question. We report a deep dive analysis of six of the modeled aspects of pilot performance: pilot error, workload, multi-tasking, situation awareness, pilot-automation interaction, and roles & responsibility, focusing attention on the nature and findings of the several models within each section, including the extent and quality of validation and verification. We then describe in detail, several ***model architectures****, that appeared in more than one of the modeling efforts in our deep dives. We then present overall conclusions and recommendations in a final section. We emphasize the importance of continuing to validate the models, in particular those that accommodate more than one aspect of pilot performance within their architecture.*

1. Introduction

NextGen operations are associated with a variety of changes to the national airspace system (NAS) including changes to the allocation of roles and responsibilities among operators and automation, the use of new technologies and automation, additional information presented on the flight deck, and the entire concept of operations (ConOps). In the transition to NextGen airspace, aviation and air operations designers need to consider the implications of design or system changes on human performance and the potential for error. To ensure continued safety of the NAS, it will be necessary for researchers to evaluate design concepts and potential NextGen scenarios well before implementation. One approach for such evaluations is through human performance modeling. Human performance models (HPMs) provide effective tools for predicting and evaluating operator performance in systems. HPMs offer significant advantages over empirical, human-in-the-loop testing in that (1) they allow detailed analyses of systems that have not yet been built, (2) they offer great flexibility for extensive data collection, (3) they do not require experimental participants, and thus can offer cost and time savings.

HPMs differ in their ability to predict performance and safety with NextGen procedures, equipment and ConOps. Models also vary in terms of how they approach human performance (e.g., some focus on cognitive processing, others focus on discrete tasks performed by a human, while others consider perceptual processes), and in terms of their associated validation efforts. The objectives of this research effort were to support the Federal Aviation Administration (FAA) in identifying HPMs that are appropriate for predicting pilot performance in NextGen operations, to provide guidance on how to evaluate the quality of different models, and to identify gaps in pilot performance modeling research, that could guide future research opportunities. This research effort is intended to help the FAA *evaluate* pilot modeling efforts and *select* the appropriate tools for future modeling efforts to predict pilot performance in NextGen operations.

1.1 Overview

The project consisted of eight primary tasks:

Task 1: Research the available HPMs used for flight deck aviation.

To address Task 1, the team reviewed literature to identify models that have been used to evaluate pilot performance. A spreadsheet was developed to summarize the relevant pilot performance modeling efforts and empirical validation studies of pilot model predictions.

Task 2: a. Develop a classification scheme for HPM evaluation and apply this classification scheme to the available HPMs identified in Task 1.

b. Characterize how criteria can be assessed to define research requirements and to evaluate research products by identifying and enumerating those that predict specific NextGen-relevant aspects of pilot performance, and those that have included a validation component.

For Task 2, the team identified a preliminary set of coding criteria or features to characterize models. Through an iterative process of reviewing papers that summarize modeling efforts, applying the coding, and identifying problems (categories that were not relevant, issues that were not captured in the criteria), the team developed a final set of coding features and criteria. Based on these coding criteria, an assessment of the state-of-the-art of modeling capabilities was made by determining the degree to which validated models exist for each pilot performance parameter.

Task 3: Develop an approach for performing a deep dive assessment of specific pilot performance models.

We performed a preliminary deep dive assessment of the error models by reviewing the relevant models, identifying the issues that they address, and assessing the extent to which these models were validated.

Task 4: Examine and summarize the empirical database of existing pilot performance models to identify tools and modeling architectures and aspects of pilot performance that they address.

In this task, we identified the modeling architectures used to create the different pilot performance models, and made note of the aspects of pilot performance that these models were used to address.

Task 5: Determine the criteria for model verification (e.g. process verification, procedural verification), and review the database of pilot models to identify which models have been verified and the extent to which they have been verified.

We identified a set of criteria by which to characterize model verification and assessed the models that were included in the deep dive reviews in terms of their degree of verification or validation.

Task 6: Evaluate the pilot performance model validation efforts to provide assessments of validation quality and facilitate comparisons across validation efforts.

We evaluated the modeling efforts (included in the deep dive reviews) in terms of the extent to which they had been verified and validated.

Task 7: Review aspects of pilot performance addressed by modeling efforts. In particular, these are six performance aspects identified in consultation with the FAA to be of the highest priority for NextGen: pilot error, workload, multi-tasking, situation awareness, pilot-automation interaction, and roles & responsibility. Perform a qualitative analysis to describe and characterize these efforts.

We reviewed aspects of pilot performance addressed by modeling efforts, and performed a qualitative analyses, or deep dive reviews, of the six aspects of pilot performance.

Task 8: Based on the modeling efforts reviewed and criteria developed, provide recommendations of validated models that are particularly relevant to NextGen. Identify NextGen research gaps based on insufficient validation efforts as determined from the review.

We provide a set of general recommendations regarding the use of models and needs for future research.

In this report we first overview NextGen operations and modeling requirements. Then in Section 2, we describe our methods for identifying and classifying pilot models, discuss the issue of model validation, and present the overview of all pilot models identified. In Section 3, we present the deep dive into the six model aspects that were identified as highest priority for NextGen. In Section 4, we review the architecture of models that were used in multiple applications. Finally in Section 5, we present conclusions regarding the nature of the existing models, their status of validation, and recommendations for future research.

1.2 NextGen Operations and Modeling Requirements

1.2.1 Characterizing NextGen Operations

One of the main goals in this effort was to help the FAA in selecting models to predict pilot performance in NextGen operations; a suite of operations and technology intended to increase the efficiency of the future airspace, with no loss in the current high level of safety. In 2009, several members of the current research team performed an extensive analysis of NextGen operations and identified changes to ConOps compared with current day operations (see Gore et al., 2009 for details). A short summary of the key issues is provided below (from Gore et al., 2009, pp 28-30). The 2009 report addressed super density operations; this concept is particularly relevant for the FAA in evaluating pilot performance, as it provides high workload conditions for pilots in NextGen operations.

The following subsection identifies changes in technologies and procedures associated with NextGen operations. These are the types of changes that we believe the FAA will want to evaluate using human performance modeling. This section summarizes by identifying relevant NextGen human performance concerns, and how modeling can be used to address them.

"Super density operations" (SDO) is a term with several meanings within the air transportation research domain. It is one of eight key capabilities identified by the Joint Planning and Development Office that define the proposed Next Generation Air Transportation System vision (JPDO, 2007). SDO also defines a research focus area for NASA's NextGen Air Traffic Management (ATM) Airspace Project. SDO has also become a fairly generic term to describe the uniquely constrained and complex challenge of operations at, and near, major airports and terminal area airspace. The characteristics of SDO operations which set it apart from the other air transportation research domains reflect the density and complexity of the operations, the relative immaturity of research to date to address this complexity, the degree to which weather cannot be easily avoided, and the constraints applied by environmental considerations which are not as prevalent in the enroute operational sphere (which has been more fully studied, and is a more mature research discipline).

The key to NextGen SDO and what makes SDO so important is that it will enable increased traffic flows at congested airports without the need to construct new runways, which are very expensive, or even new airports at busy metroplexes, which are even more expensive and may be impossible due to the lack of available land.

The following are some of the SDO concepts and technologies intended to make better use of the scarce resources – runways and airspace – at the U.S.'s largest airports:

- ***Closely spaced aircraft*** – separations reduced to much less than today's standards due to better resolution of aircraft positions and better information available on flight decks that will help avoid midair collisions and wake vortex encounters (see next bullet).
- ***Wake vortex information*** – since current separations are conservative, due in part to the need to avoid wake vortices, reduced separation will require real-time data on wake vortex generation and dispersion. This will require sensors and models to measure and predict wake trails.
- ***Paired aircraft*** - a "daisy chain" of paired leader-follower aircraft, especially on arrival and approach. With more traffic information available on flight decks, airport operations can be conducted in almost any weather condition as if it were a clear visual meteorological conditions (VMC) day, where, in current-day operations, airliners follow each other to the landing runways. Pairing allows for closer traffic spacing and a smoother arrival flow with less workload for air traffic controllers, managers, and pilots.
- ***Very closely spaced parallel approaches (VCSPA)*** – this might involve paired aircraft, or it might involve groups of three aircraft. These three would be very closely spaced (e.g., 750 feet lateral separation), and the following group of three would be about 2 minutes behind them. This procedure, again, makes better use of the spacing in current-day operations.
- ***Trajectory based operations*** – aircraft will be assigned four-dimensional (4D) trajectories (3 spatial dimensions plus time) and expected to meet path and time requirements. Several NextGen concept developers cautioned that there is much uncertainty in how rigid these requirements will be. The 4D "tunnel" might actually be quite large, and it is currently not known what time precisions will be required.
- ***Weather information*** – to help ensure that trajectories are achievable, real-time weather data will be provided to ATM and pilots. Since weather is a major factor in reducing airport departure and arrival rates, making real-time weather data and information available will allow for anticipation of weather-related delays and the application of suitable contingency plans in a timely and more efficient manner.

- ***Continuous descents & ascents*** – for environmental and economic reasons, leveling-off flight will be minimized. Level-offs will be limited to the cruise phase. The more time spent by aircraft at low altitudes, the more fuel burned by those aircraft.
- ***Datalink communication with ATM*** – rather than voice communication, NextGen communication will be electronic, visual, and text-based (like instant messaging or email). The benefit to this technology is that complex clearances, such as directions for paired approaches or 4D paths, can be communicated more quickly and accurately, and then easily loaded into the aircraft flight management system (FMS). The downside to datalink clearances is the added visual workload for the pilot, and the fact that mistakes can be more easily over-looked and propagated. Therefore, error (e.g., keyboard entry) and logic (e.g., is there a more efficient path?) checking seem essential to take full advantage of this capability.
- ***Uplinked taxi information*** – taxi clearances will be provided via datalink before the aircraft lands, thus minimizing the time spent between the runway and parking gate.
- ***Equivalent visual operations*** – electronically-generated out-the-window view (with synthetic or enhanced vision displays and real-time sensing capabilities) will potentially reduce decision height, and hence better preserve landing capabilities in low visibility.
- ***Mixed equipage operations*** – many different aircraft with many different capabilities will (potentially) mean prioritized flights, perhaps segmented airspace or timeslots. This has the potential for blunders into airspace, and pilot or ATM errors regarding aircraft capabilities. In current-day operations, aircraft without specified capabilities are not allowed into the most congested airspace (i.e., Category B airspace), so keeping less capable aircraft out of metroplex airspace should improve efficiency. The "flip side of this coin," though, is that when insufficiently equipped aircraft blunder into more tightly controlled airspace, it is likely that such blunders may cause major delays, inefficiencies, and other impacts.
- ***Performance based services*** – In the evolving and future (e.g., NextGen) airspace, there are anticipated to be a larger number of different airplane equipage capabilities, such as those enabling self-separation. Similarly, current operations accommodate different levels of **required navigational performance** (RNP) such that greater precision can enable more economical operations and trajectories.
- ***Self-separation*** – Aircraft with particular equipment (e.g., the future equivalent of automatic dependent surveillance-broadcast [ADS-B] and a cockpit display of traffic information [CDTI]), will be able to carry out tactical maneuvers to maintain separation from other traffic, in the absence of positive guidance from ATC.
- ***Metroplexes*** – capacity increases will be met by groups of airports that effectively function as one large airport (e.g., Newark, LaGuardia, and JFK; or San Francisco, San Jose, and Moffet Field). This may mean more complex traffic patterns into and out of the airports, but more efficient operations overall.
- ***Net-centric operations*** – NextGen will rely heavily on computerized information systems (e.g., route planning capabilities for 4D trajectories, digital maps, pilot-ATM and pilot-pilot communication, replanning and rerouting capabilities, synthetic vision generation, weather and wake vortex updating and visualization) and the timely exchange of information. This need implies that computing power on the flight deck will need to be much greater than current standards, and that cyber security (i.e., network security) is extremely important.

Most of these NextGen concepts will require that additional information is displayed on the flight deck. While the form of this information (what it will look like) and its location (if it will be integrated into existing displays, or presented on a new display) are uncertain, the project team

assumed that the following information sources are available on the NextGen (circa 2025) flight deck:

- Datalink text and possibly graphical messages
- Wake vortex (WV) information
- Integrated weather information (Note: this is currently on the Navigation (NAV) Display in the Boeing 777.)
- Vertical situation display (VSD)
- Location of, and separation from, other aircraft, with particular focus on the lead aircraft (providing coverage beyond the traffic collision avoidance system, TCAS, display) (Note: this is currently included on the NAV display of the B-777.)
- Equivalent visual operations (EVO) using synthetic vision system (SVS) or enhanced vision system (EVS) information located on a head-up display (HUD) or other flight deck display
- Uplinked taxi clearance (provided via datalink and/or a dynamic airport surface map or head-up display)
- Runway Awareness and Advisory System (RAAS)
- Merging and Spacing
- Electronic Flight Bag (EFB)
- Cockpit Display of Traffic Information (CDTI)
- Higher levels of flight deck automation, particularly within the FMS, to assist decision making

1.2.2 Modeling Concerns Related to NextGen

As the above lists indicate, NextGen operations are associated with a variety of changes to the NAS. These changes affect the allocation of responsibilities, the use of new technologies and automation, additional information presented on the flight deck, and the entire ConOps. A recent integration of human-factors-relevant NextGen research identified nine main human performance factors to consider in NextGen operations: attention allocation; decision making and mental modeling; communication; memory; workload; interaction with automation, decision support tools, and displays; potential errors and recovery from human errors or system failures; organizational factors; and selection, certification, and training (Lee, Sheridan, Poage, Martin, Jobe, & Cabrall, 2010). In this report we review and classify existing models to assess the extent that HPMs are capable of addressing each of these NextGen-relevant human performance parameters (note, however, we focus on the first seven issues listed above, as organizational factors, selection, certification, and training are beyond the scope of pilot performance models).

Modeling provides a capability to examine potential changes to a system design before these changes are introduced. This allows analysts to predict the effects on operator and system performance, and to identify and remediate problems. However, models vary in their abilities and their approach to modeling these human performance factors (e.g., some focus on cognitive processing, others focus on discrete tasks performed by a human, while others consider perceptual processes), and in terms of their associated validation efforts. The challenge for the FAA is in evaluating the variety of pilot performance models and modeling tools to identify those that are best suited to addressing these NextGen human factors issues.

2. Identification and Classification of Models and Model Validity

2.1 Identify Available Pilot Performance Models

The current effort comprehensively reviewed the current literature to identify all relevant HPMs that have been applied to assess pilot performance and which have, or can be, extended to evaluate NextGen operations.

2.1.1 Methods

The team identified a wide variety of sources that described aircraft pilot modeling and validation efforts for the initial pass. These sources included professional journals, conference proceedings, and reports from specialist working groups. Table 2.1 lists the sources that were included in the search. A relevant set of keywords was developed and used to search among the sources; these are provided in Table 2.2. The focus of the initial data gathering was on ***pilot performance models***. More general models of human performance or human error, and specific models of air traffic management and unmanned air vehicle control were excluded. Further, models that were based on pilot performance during training, and used for personnel selection, were considered not relevant, and were also excluded from this data set.

Table 2.1: Sources of Articles Describing Pilot Modeling Efforts

Source	***Location***
A Review of Human Performance Models for the Prediction of Pilot Error	http://humanfactors.arc.nasa.gov/ihi/hcsl/publications/HumanErrorModels.pdf
Advanced Simulation Technologies Conference	http://www.scs.org/confernc/astc/astc04/cfp/astc04.htm
Advanced Simulation Technologies Conference	http://www.scs.org/confarchive
Agents and Artificial Intelligence	http://www.icaart.org/
American Institute of Aeronautics and Astronautics (AIAA) Proceedings	http://www.aiaa.org/content.cfm?pageid=406
An HCI bibliography site	http://hcibib.org/bibtoc.cgi?abstracts=true&file=bibdata/HFES10*
Annual Conference on Manual Control	CDs – Conference proceedings
Applied human factors and ergonomics (AHFE)	http://www.ahfe2012.org/
Artificial intelligence	http://www.aaai.org/Conferences/conferences.php
Aviation Space and Environmental Medicine	www.ingentaconnect.com
BRIMS Proceedings	http://brimsconference.org/
Citeseer	www.citeseerx.ist.psu.edu/viewdoc/

Source	***Location***
Cognitive Engineering and Decision Making	www.ingentaconnect.com
Cognitive Engineering in the Aviation Domain	Book (Sarter & Amalberti, Eds., 2000)
Digital Aviation Systems (DAS) Conference Proceedings	http://ieeexplore.ieee.org/xpl/conhome.jsp?punumber=1000202
Ergonomics in Design	www.ingentaconnect.com
HCI Aero Conference Proceedings	Subscriber database
Human Computer Interaction International (HCII)	http://www.hcii2011.org/ (2011) http://www.hci-international.org/index.php?module=conference&MMN_position=4:4 (previous years)
Human Factors and Ergonomics Society Conference Proceedings	www.ingentaconnect.com
Human Factors Journal (HFJ)	www.ingentaconnect.com
Human Performance Models in Aviation	Book (Foyle & Hooey, Eds., 2008)
Human Performance Situation Awareness and Automation (HPSAA II)	CDs – Conference proceedings
IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC)	Subscriber database
International Journal of Applied Aviation Studies	http://www.faa.gov/about/office_org/headquarters_offices/arc/programs/academy/journal/
International Journal of Aviation Psychology (IJAP)	www.informaworld.com
International Journal of Human Factors Modeling and Simulation	Subscriber database
International Society for Human Simulation	http://www.societyhumansimulation.org/content/about-ishs
International Symposium on Aviation Psychology (ISAP)	CDs – Conference proceedings
Model Based Analysis of Human Errors During Aircraft Cockpit System Design (European Union Project)	http://www.human.aero *(many articles were available on www.mendeley.com)*

Source	*Location*
NASA Human Centered Systems Lab page – publications (contractor reports that went into the Foyle & Hooey (2008) book *Human Performance Modeling in Aviation*)	http://hsi.arc.nasa.gov/groups/HCSL/publications.html#HPMPubs
NASA Technical Reports Server	http://ntrs.nasa.gov
National Research Council	http://www.nap.edu/catalog.php?record_id=12169
NATO human behavioral modeling	ftp.rta.nato.int/public//PubFullText/RTO/MP/...///MP-HFM-202-P08.doc
Proceedings of the World Aviation Conference	Identified as references in other articles – did not find a website or document.
Sim Solutions Conference Proceedings	CDs – Conference proceedings
SOAR Reports	http://sitemaker.umich.edu/soar/home
Theoretical Issues in Ergonomic Science	http://www.tandf.co.uk/journals/ttie Subscriber database
Web sources for HOS and SAMPLE	Google search
Winter Simulation Conference	CDs – Conference proceedings

Table 2.2: Keywords used in the Data Gathering Effort

Keyword	*AND*	*Other Keyword*
Aerospace		Model / Simulation
Aviation		Model / Simulation
Pilot		Model / Simulation
Pilot		Workload / Error / Errors / Performance / Models / Modeling / Modeling / Behavior model
Predicting		Pilot performance

2.1.2 Results

The team compiled a database of the existing pilot performance models and relevant validation efforts. The initial pass revealed 449 potentially relevant articles. A second pass review was conducted, which consisted of identifying 1) duplicate articles, 2) nearly identical articles (e.g., same set of authors presenting similar research at two different conferences), 3) models that addressed air traffic management (ATM) rather than pilot performance, and 4) articles that focused on empirical studies of pilot performance without including or addressing a

corresponding pilot model. When this second review was completed, the list of articles was reduced to 187. The final list of articles, with full references, is provided in Appendix B.

2.2 Classify Pilot Performance Models

2.2.1 Develop Classification Scheme

The team identified a preliminary set of coding criteria or features to characterize models. Through the iterative process of reviewing papers that summarize modeling efforts, applying the coding, and identifying problems (categories that were not relevant, issues that were not captured in the criteria), the team developed a final set of coding features and criteria, presented in Table 2.3.

Model features used in this evaluation belong to one of three 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 are features we refer to as ***"criteria"***, whose rating for any particular model ***does*** implicitly or explicitly suggest greater or lesser value or relevance to the FAA. For example, all other factors being equal, a validated model is more valuable than an unvalidated one. This second general class of evaluative features can be further subdivided. One set of evaluative features (class B) is 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. These are features that will be applied to the relevant models in this final report. A third set of evaluative features (class C) includes those features that are important in assessing the overall value or ***utility of the model*** to the FAA, but are not related to validation and, in general, are features such as usability and software support that the current research team cannot readily evaluate within the scope of this project. Regardless, these are criteria that the FAA may want to consider when selecting models. The list of features, of all three classes was based upon a CSERIAC report (Wickens, Vincow, Schopper, & Lincoln, 1997). 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. While class C factors were not evaluated systematically for each modeling effort, this report does identify relevant information on accessibility and support in Section 4.

Altogether 12 features were defined to characterize each separate research paper that examines a pilot performance model, with each categorical feature having a number of different levels, and shown in Table 2.3. Features 1-3 are descriptive (class A), and features 4-9 are evaluative (class B) features, and features 10-13 are additional or utility (class C) features. These features are described below.

Table 2.3: Pilot Model Classification Features: Factors, Codes, and Meanings

Evaluation Factors	*Code*	*Meaning*
Class A: Descriptive Features		
1. Type of modeling effort		
	Si	Simulation
	An	Analytical
	Re	Regression equation
	Q	Qualitative
2. Aspect of pilot performance modeled [not mutually exclusive] Those aspects indicated with a * are selected for a deep dive appearing in Section 3.		
*	A	Automation interaction
	C	Communications
	D	Decision or judgment
*	E	Error
	MC	Manual Control
*	MT	Multi-task (task sequencing, multiple resources)
	P	Procedures
*	SA	Situation awareness
	SD	Spatial disorientation
	V	Vision, visual attention, scanning
*	W	Workload
*	R&R	Roles & Responsibilities
3. Model name (if appropriate)		
Class B: Evaluative Features		
4. Empirical data available		
	Y	Yes
	N	No

Evaluation Factors	***Code***	***Meaning***
5. Validation measure		
	C	Correlation
	Q	Qualitative
	O	Other
6. Correlation parameters		
	Value	(0.XX)
	Sample size	Number of data points (not number of subjects)
7. Correlation methods		
	BC	Between conditions (collapsed across subjects)
	SAC	Between subjects and conditions
	CWS	Between conditions and within subjects
	Other	Not specified, or another technique used
8. Population		
	TP	Transport Pilots (includes commercial, corporate, and cargo transport)
	NP	Not pilots
	FP	Fighter pilots
	OP	Other pilots (e.g., general aviation, helicopter)
9. Test bed		
	L	Laboratory
	PC	PC flight simulator (mouse / joystick)
	FS	Flight simulator (pedals / yoke)
	AP	Airplane
Class C: Additional Criteria		
10. Cognitive plausibility		
11. Usability		
12. Availability		

Class A: Descriptive Features

Feature 1: ***Type of modeling effort***. The modeling papers in the analysis were classified according to model type. The model types included simulation, analytic, regression, or qualitative. Simulation models produce differing results due to performance distributions and built-in randomness. Analytic models are algorithms or equations that yield a single set of results for given input parameters. Regression equations are quantitative explanations of performance, and qualitative models predict non-numeric aspects of performance (e.g., better or worse). A qualitative model is simply a description of processes that contains no component that will generate numerical output.

Feature 2: ***Aspect of pilot performance modeled***. After reviewing the available models, the **aspect** categorization appeared to encompass all models, and each term provided a useful aviation-relevant description that could also be associated with keyword searches. It was important that these categories not be 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). Indeed we found several models that addressed multiple aspects. We note that the * aspects within the left hand column are those selected for a more intense deep dive, following consultation with our sponsors at the FAA midway through the project.

Feature 3: ***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. These architectures are reviewed in detail in Section 4 of the report.

Class B: Evaluative Features

Feature 4: ***Empirical data available***. Here we determined if empirical human performance data were reported in the paper, *that could be directly employed (or were employed) to evaluate the model.* Thus in some cases a paper reported data, and a model, but the data were not to be predicted by the model. Here the classification would be "N". A "Y" rating was reserved for those cases where data were available to evaluate the quality of model predictions.

Feature 5: ***Validation measure***. The correlation measure is self-evident. A qualitative validation might include a statement like: "the pattern of errors predicted by the model was quite similar to that shown by the pilots." In some cases, the model predictions and data were reported in such a way that a quantitative estimate could be derived by the reader of this report. Further details regarding the use of correlation for validation are included in Section 2.2.3 and Appendix A. Issues related to defining validation are discussed in Section 2.2.3 below.

Feature 6: ***Correlation parameters***. These numerical terms describe the results of a validation study, through the value of the correlation and the sample size. The sample size does not refer to the number of participants in the study, but rather the number of conditions across which the bivariate point of a [model prediction – PITL data point] could be determined (e.g., 3 conditions of weather; n=3).

Feature 7: ***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. For example, if a validation effort examines error prediction in four different NextGen procedures, the model would predict error rate for each of these four, and the validation experiment would generate actual pilot errors observed for the four procedure conditions. A correlation (N=4) of predicted versus obtained error rates could be computed. This aspect of correlation validation is described as "between conditions" (BC). This measure is desirable because it captures the variance accounted for by the model (r-squared) independent of N, so long as N >2). If the bivariate data set from which the correlation is obtained includes not only differences between conditions, but also the observable data from different pilots, this is described as subjects and conditions (SAC). If a set of different conditions is validated multiple times, once for each subject pilot (e.g., generating N scatter plots & correlations, where N is the number of pilots), then the code CWS (conditions within subjects) is applied.

Features 8 and 9 (***Population*** and ***Test-bed***) are both self-evident in their descriptive labels, but each can 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 relevance to the FAA than studies employing less-representative populations or test-beds.

Class C: Additional Criteria

Feature 10. ***Cognitive plausibility*** describes the extent to which parameters and variables in the model replicate, mimic or are linked to processes known from human psychology. Making this type of determination requires understanding how the model predicts human performance, and the theoretical underpinnings of these predictions. The evaluator must decide if the model provides a reasonable interpretation and instantiation of this theory.

Feature 11. ***Usability***: This feature describes many aspects of the usability of the model: to what extent are the parameters to be entered well explained in a model manual or guide; to what extent developers have produced an easy-to-use interface. These types of evaluations should be made by reviewing the model, attempting to enter data, and attempting to make sense of the commands. General evaluative rules are that more usable models offer graphical user interfaces and provide on-line help and documentation such as user manuals. Less-usable models require programming expertise and do not offer readily understood documentation. The ability to attend training is another important factor to consider. Modeling tools are typically complex (i.e., they are not "walk up and use" systems). A modeling tool for which training is readily available might be a better choice than one for which no training is offered.

Feature 12. ***Availability***. This describes the extent to which the model can be applied from a readily available product (i.e., commercially available or government product). Some products are available online, but these are frequently provided without extensive user support. Other issues to consider include the year the software was developed (more recently developed products are typically better than those that have not been changed in a couple of decades), and the type of support that is available to users.

A table that presents a qualitative comparison of the extent of verification and validation across modeling efforts that were reviewed in the deep dive analysis was generated and is presented in Section 3.6 to enable an "at-a-glance" review of the model set considered.

It would also be possible for readers of this report, to carry out a more extensive evaluation of different models based upon the additional criteria for relevance to NextGen. This score could be weighted, based on differences in importance of different features, or unweighted (a simple summation). As noted however, it will not be our intention to score models on Class C features. We do intend, however, to elaborate on the descriptions of these criteria, to assist the FAA in evaluating models according to these features.

2.2.2 Classify Models

Each of the 187 articles identified in our search was reviewed and coded according to the 9 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 or more validation experiments of a single model or if the paper presented two or more 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.

The list of sources used in this evaluation is provided in Appendix B. Further, all sources, and the ratings associated with them, are included in a spreadsheet entitled: FAA_HPM_Annex11_Deliverable.xls. This is available from the first authors upon request (cwickens@alionscience.com, or asebok@alionscience.com).

2.2.3 Define Validation

The focus of this research on model validation requires operational definitions of terms. We consider the term **"model"** here to be defined as in two National Research Council books:

> "Human….behavior…represented by computational formulas, programs or simulations" (Pew & Mavor, 1998) and
>
> "A representation or description of all or part of an object or process" [p11] (Elkind et al., 1990).

In particular, the second of these references goes on to describe the functional use of models in human factors to be to "answer questions about the ability of the human to function satisfactorily in the system, and the ability of the system to achieve the objectives for which it is being designed" [pp3-4]. While Elkind et al. (1990) place their emphasis on simulation or Monte-Carlo models which can be run, and include an aspect of randomness or distributions in performance parameters, we expand our search here to include that which Pew and Mavor (1998) refer to as computational formulas, and Wickens, Vincow, Schopper and Lincoln (1997) describe as "analytic models". Analytic models are equations that predict performance based on certain

input parameters, but they do not include an aspect of randomness and do not require repeated iteration of model runs to generate that randomness or expected distribution of performance parameters. The same input conditions will provide identical results.

The operational definition we use for "**validation**" is equally important, and here we rely on Pew and Mavor (1998) to define validation as showing "that the model *accurately represents behavior* in the real world under at least some conditions" [p4, italics ours]. As such, this definition closely reflects what Leiden and Best (2008) describe as "**results validation**," or "the ability of the model to provide sound predictions within certain bounds" [p278].

In contrast to validation, **verification** involves providing "proof that the model actually runs and meets the design specifications" [Pew & Mavor, 1998, p278], or that the model "performs as intended by the modeler" [Leiden & Best 2008, p278]. This may include sensitivity tests, in which model outputs are generated across a range of different parameters. There is in fact somewhat of a fuzzy boundary between verification and qualitative validation. However in general we refer to qualitative validation as those judgments made by personnel other than the modeler developers themselves (e.g., pilot SMEs).

While verification is typically necessary for the success of a model, our focus in this report is more directly targeted at ***results validation***. We are concerned with addressing questions such as:

- Does the model predict variance of real-world pilot behavior?
- Across what conditions are those predictions made (e.g., all of the conditions predicted by the model, or a limited subset)? and
- What is the metric by which the degree of prediction between model prediction and pilot performance data is expressed? Here of course a distinction can be made between whether such a prediction is available (and this, by definition, requires reporting of some pilot-in-the-loop data) and **how successful that prediction is.** Determining the latter requires the authors of this report to provide some evaluation of the metric used, and its predictive value (e.g., correlation level between model-predicted and obtained data) in a particular evaluation study. Model validation evaluation is described in more detail in Appendix A.

The expressed desirability of the product moment correlation, r, when computed between conditions (BC, for feature 7, Correlation Method, above) results from the fact that this is the best measure of the extent to which variance between conditions predicted by the model are actually observed in performance. The presentation of the scatter plot upon which the correlation is based, in addition to the r value, is important because it allows the reader and model-user to visualize which conditions may be over or under predicted by a model, or whether a high correlation value results because of the contribution of only a single outlier (See Appendix A).

The reporting of the significance level of the correlation, while sometimes desirable, is not necessary when N is small. The low N will often "guarantee" non-significance, when the pattern of prediction evident in the scatter plot and r value may be extremely meaningful. For example,

consider a validation with only four conditions (N=4) defined by 2 levels of NextGen technology by 2 levels of workload that yields an r of 0.80. While this may not lead to a "significant" correlation, its high value certainly suggests that the model is effective in predicting variance between conditions.

Finally, we wish to emphasize the important distinction between saying a particular modeling effort **has not been validated** versus a model is "**invalid**". The "not yet validated" situation occurs for a variety of reasons, including the fact that sometimes, particularly for systems under development such as NextGen, the necessary technology to accomplish a validation has not yet been implemented in a flight simulation. Even though the model predictions cannot be validated, they are still important and useful. In contrast, to say that a model is "invalid" (a characterization that we do not use in this report), would imply that over repeated validation efforts, its correlation with empirical data is near 0.

2.2.4 Assess overall Validation Efforts across all Model Aspects

The number of models that could be fit into each category (e.g., manual control, error) was tallied based on the assigned pilot performance modeled category. A given model validation was sorted into more than one category 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). Table 2.4 lists the frequency count within each of these categories, along with a second column that reports the number of studies that were considered to have validation data.

Table 2.4: Types of Models and Validation efforts – Total Numbers and Percentages.

Performance Modeled	***Number***	***Number Validated***	***Percent Validated***
Error	17	9	53
Workload	14	See below	
Multi-task (& task management)	19	See below	
*Workload & Multi-task**	*(33)*	*14*	*42*
Situation awareness	15	3	20
Pilot-automation interaction	16	8	50
Communications	7	0	0
Decision-making	22	6	27
Fatigue	3	2	67
Manual control	30	20	67
Procedural	26	5	19
Roles & Responsibilities	12	3	25
Spatial disorientation	1	0	0
Visual	49	31	63
SUM	**231**	**101**	**43%**

* Workload and multi-task were pooled because of the close overlap between the two constructs, and occasional difficulties in discerning whether workload or multi-tasking predictions were validated. See Section 3.2 for further explanation.

The overall statistics revealed that of the 231 samples, 43% were validated (calculating the unweighted average). The number of samples here was greater than the number of studies, because some models were coded into multiple categories. A few noteworthy aspects of the data concern the 0% validation rate of the 7 communications models and the relatively low (25% or less) validation rate of procedural models and models of situation awareness and roles and responsibilities.

3. DEEP DIVE ANALYSIS OF SIX MODEL ASPECTS

The following "deep dive" sections include reviews of papers that were analyzed in detail and included in the spreadsheet. These are the references that are included in the summary tables throughout this document (e.g., Table 2.4, Table 3.6, Table 4.1), and – for each individual deep dive – they are identified in a reference list at the end of that particular section.

In addition, each deep dive also includes references that were ***not*** included in the project spreadsheet or summary statistics. These were excluded from the analysis for various reasons (e.g., they reported an essentially duplicate effort summarized in another [included] paper, or the paper did not actually include a model of pilot performance), but they were relevant to, and therefore referenced in, the deep dive sections. These are not included in the reference list at the end of the section, they are not in the spreadsheet, nor are they included in the summary statistics in this document. They are, however, presented in the full reference list at the end of the document.

3.1 Models of Pilot Error

3.1.1 Introduction

This section describes a "deep dive" review of the pilot performance models that specifically addressed error to identify common themes among those modeling efforts. While this review specifically evaluated human error models that have been applied to pilot performance, it should be noted that Leiden et al. (2001) reviewed HPMs that address the broader category of human error. These tools could potentially be used to model pilot error. It should be noted that the type of error that can be modeled by the respective piece of software does depend on its underlying psychological principle (if existent).

In reviewing the variety of models to address pilot error, it is important to consider how errors occur and how they are modeled. For example, Leiden et al. (2001) identified a set of error taxonomies, including: situation awareness errors (errors in detecting, comprehending, or predicting based on data), models of internal human malfunctions (errors in terms of operator factors such as goal setting or strategy selection), models of unsafe acts (Rasmussen's skill-based, rule-based, and knowledge-based errors), and information processing models (which simulate human cognitive processes of detection, recognition, decision making and action selection).

3.1.2 Human Reliability Analysis

A set of 3 papers by the Salmon and Stanton group (Salmon et al., 2002; Stanton et al., 2003; and Salmon et al., 2003) all evaluated different means of classifying pilot error, with a focus on the Systematic Human Error Reduction and Prediction Approach (SHERPA). The validation is 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 pilot subject matter experts (SMEs). Both predictions were made within an auto-land flight scenario. Note that these were *not errors actually observed* as no simulation was flown, so that it constituted a qualitative validation. 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 indicate 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. The paper does not offer quantitative predictions (e.g., of the relative likelihood of different kinds of errors, or the estimated likelihood of error in different scenarios) nor are validation data provided.

3.1.3 Procedural Risk Models

A set of four papers developed at NLR (Netherlands Aerospace Research Lab), by Blom, Stroeve and their colleagues was centered around one generic model – TOPAZ (Traffic Organization and Perturbation Analyzer; Stroeve, Blom, & Baaker, 2011; Stroeve, Blom & Baaker, 2009; Stroeve & Blom, 2005; Blom, Corker, Stroeve & van der Park, 2003) – and one specific application: predicting runway incursions resulting from one aircraft failing to stop at an intersection where another was on a takeoff 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). The studies report **verification**, in that 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 & 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.

Correspondingly, the emphasis on airport surface procedures and runway incursion event is quite relevant to near-term NextGen concepts. As noted, there is no validation, since there was no reporting of the actual observed frequency of such low probability events, as they might be affected by display and environmental variables.

3.1.4 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 design of a fighter-cockpit interface in predicting different forms of errors (e.g., attention, planning, motor), but the paper offers no validation.

- 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. In this database, a corpus of twelve errors committed by eighteen two-person 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). It is important to note that "the modelers were not tasked to **predict** the taxi error data set, the observed error rate, or proportion of errors for each error category. Rather, they were asked to use the data set and other available information to help build solid models of pilot behavior that would be the foundation of a set of tools with which taxi navigation errors, causes and mitigations could be explored" (Foyle & Hooey, 2008, p. 38).

- 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 taxiway 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 & 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). The authors stated that the model "created a diversity of errors" including errors of omission and commission. An error of omission occurred when the modeled pilots could not recall a required taxi turn instruction due to time-based decay, leading to a missed turn. Errors of commission, leading the modeled pilot to turn on the wrong taxiway or turn the wrong direction, occurred when the wrong taxi instruction was recalled because of interference,

similarity-based partial matching, priming, or activation noise. Lebiere et al.'s model focused on memory errors given the nature of the HITL data, but they reported that other possible sources of errors including perceptual inaccuracies, loss of SA, and pilot distraction could also be modeled in a principled approach in the ACT-R architecture if enough HITL data were available.

3.1.5 Error Generation Models

- The chapter by Deutsch and Pew (2008, in Foyle & Hooey) describes a model, the Distributed - Operator Model Architecture (D-OMAR), which employs ACT-R-like 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 addresses and describes 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). The results of the model simulation are presented in terms of a detailed narrative description of the kinds of errors, but not accompanied by quantitative validation data of error prediction against the error data provided.

- ***"Air" MIDAS*** (MIDAS 1.0) (Corker et al., 2008 in Foyle & Hooey) is a full pilot model that has received fairly extensive validation in other model categories (e.g., workload, procedures, visual attention, roles & responsibilities), and those validations will be provided in sections 3.2 and 3.6. 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. Working memory decay rate and capacity, and timing of information availability to the operators in the simulation were modeled. Error rates, performance times and workload were output from the HPM. These predicted that errors do occur as a result of increasing decay rates and reducing memory capacity, while changes in the data patterns occur as a result of changes to the time of information availability to the crew members in the simulation. This model application included both memory failures and **workload** influences. The authors reported that Air MIDAS is sensitive to memory errors including declarative memory errors (forgetting a procedure because of having too many procedures of the same type operating at a given time), memory load errors (as a result of information competing for working memory space) and updateable world representation discrepancy errors (which occur when the worldview between two operators is inconsistent). The authors describe a qualitative form of validation, noting consistency of the pattern of errors between model output and NASA-provided simulation data.

- ***The A/SA (Attention-Situation Awareness) model*** (Wickens et al., 2008, in Foyle & Hooey) focuses on modeling **visual attention**, and the memory-related loss of **situation**

awareness as factors that lead to either enhanced or degraded sense of position of where the aircraft is on the runway surface. (Hence their model will be listed also in those categories). 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. Quantitative predictions of error probability are offered, as a function of the presence or absence of cockpit display support technology, at four points along the taxiway. This provided a basis for qualitative validation, and could be compared against the actual frequency of errors with and without the technology; but this correlation was not presented in the report.

- ***The Cognitive Architecture for Safety Critical Task Simulation*** (CASCaS). A computational model has been proposed by Lüdtke et al. (2009) that is a full pilot performance model and will be described in more detail in Section 3.5. 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 type learning mechanism as the context it simulates is of pilots repeatedly using the flight management system (FMS) in programming specific procedures. As manifest in the error predictions reported in the paper, both types of errors are essentially errors of attention (failure to notice incorrect states), rather than of error generation. This paper does not contain any validation data, but instead provides a verification of model-generated predictions in using the FMS.

3.1.6 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.

- Karakawa et al. (2006) presents the ***pilot cognitive simulation (PCS)*** that models full pilot cognitive capabilities. The emphasis of this simulation model is 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. The authors reported that the model predicted better detection of errors with the enhanced display, but it is unclear how the authors compared this with the actual data (improved pilot performance in error detection).

- 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 is 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 however, 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 is 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.

3.1.7 Conclusions Regarding Pilot Error Models

Pilot error models accounted for less than ten percent of the overall pilot models identified and reviewed in this survey. Models address a wide range of error issues, including error generation, classification, detection, and recovery. Models also address underlying cognitive and perceptual factors that affect error likelihood. Models typically focus on limited aspects of pilot error. Only D-OMAR and the human reliability analysis models appeared to have a broad focus to identify or predict all types of errors, rather than those of a specific class like memory retrieval failures, attention errors or decision failures.

Of the seventeen error-model papers reviewed (those that clearly had the commitment of pilot errors as the focus, as opposed to other error generating outcomes, like high workload, or inappropriate scanning), only 9 appeared to contain data against which the models could be validated, and of these, none contained quantitative validation; that is, quantitative measures of the degree of match or "fit" between model predictions and actual error commitment, either predicting differences in overall error rate within different circumstances (e.g., with or without automation), or predicting the difference in categories of errors.

Some of these models **described** the similarity (between predictions and pilot data) in the qualitative pattern of errors, and others noted that a change in predicted error rate (brought about, for example, by technology) could parallel a corresponding change actually or likely to be observed in actual pilot performance.

Others (e.g., the TOPAZ models) provided precise quantitative model predictions, but no validation data, and still others (e.g., the SHERPA) while offering quantitative validation, used as performance measures for validation, the likelihood of error predicted by a SME, rather than the probability of error from actual pilot data.

It should be noted here that the set of five different error models in the Foyle and Hooey (2008) book did contain more precise validation data from their models reported in a second part of each chapter, but that validation was not of errors per se, but of predictions of visual scanning performance, a model aspect not targeted for a deep dive in the current effort.

In accounting for this state of affairs regarding limited validation of error models, three points should be noted.

1. ***Errors generally (and fortunately) are rare in aviation, and hence it is often hard to get quantitative validation data for low-frequency events*** (but see Wickens, Hooey, Gore, Sebok & Koenecke, 2009, for a successful attempt to do so, for perceptual errors, and Nikolic & Sarter, 2003, for successful use of ASRS data). This rarity was evident in the very low number of taxiway errors (12) committed in the simulation and hence available to the modeling teams, a number that is the numerator of a very low fraction defining taxiway error rate (i.e., number of errors divided by number of opportunities for errors – intersections crossed). 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 than 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.

2. ***Errors have multiple internal (pilot-cognition) causes***, as reflected by the diversity of processes modeled in the reports above (these processes, like memory, attention, decision-making, and SA were **boldfaced)**. Generally a particular model tackles only a small sampling of these processes (the full pilot models like 'Air' MIDAS and SHERPA are the exception), and hence will be incapable of predicting differences across a more diverse set of errors. There are several classes of models, contained in table 2.3, whose validation does not predict errors per-se, but rather, address operating conditions or precursors that are known to produce errors, such as high workload, conditions inviting low situation awareness, or complex procedures.

3. Nevertheless, in some cases, ***there appears to be an opportunity for model developers to apply their validation to this valuable database, the corpus of high-fidelity simulation taxi navigation errors.*** The fact that this validation has not been done might be attributable to the added financial and time resources required, and the restriction of funding to support it, particularly at the end of a long project in which considerable funds were expended simply to understand and model the complex scenarios and generate model predictions. This suggests that modeling efforts should consist of iterative model develop-validate cycles, with early efforts taking the form of verification, moving to qualitative validation, and ending with quantitative comparisons to HITL data.

The error models described here are clearly applicable to NextGen operations. Based on the analysis of NextGen operations, the primary differences compared with current day operations are as follows: the pilot's role will shift from a manual controller of the aircraft and automation to a monitor of diverse automation systems. Further, responsibility for maintaining separation from lead aircraft is expected to shift (at least partially) to the flight deck. Pilots will be supported in this task by additional automated systems and information displays, along with another layer of health information regarding these systems. Finally, routine communication between pilots and air traffic control will be performed by datalink, rather than verbal, radio-based communications. These tasks are expected to make the pilot potentially more vulnerable to automation-related problems (e.g., complacency, becoming out of the loop, manual skill degradation) and being placed in potential visual overload conditions. However, given the consistency of the human operator, the basic human error mechanisms are expected to remain quite similar to current-day operations, so modeling tools are expected to be relevant to NextGen operations. It is critical that these models be validated in conditions that are representative of NextGen ConOps.

3.1.8 References for Pilot Error Model Data

Blom, H., Corker, K., Stroeve, S., & Van Der Park, M. (2003). Study on the integration of Air-MIDAS and TOPAZ (NLR-CR-2003). San Jose, CA: NASA/ATAC Corp.

Byrne, M.D., Kirlik, A., & Fleetwood, M.D. (2008). An ACT-R approach to closing the loop on computational cognitive modeling. Chapter 5 in D.C. Foyle & B.L. Hooey (Eds.) *Human*

Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 77 - 104.

Corker, K.M., Muraoka, K., Verma, S., Jadhav, A., & Gore, B.F. (2008). Air MIDAS: A Closed-Loop Model Framework. Chapter 7 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 145-182.

Deutsch, S.E., & Pew, R.W. (2008). D-OMAR: An architecture for modeling multitask behaviors. Chapter 8 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group, pp. 183-212.

Elkind, J.I., Card, S.K., Hochberg, J., & Huey, B.M. (1990). *Human Performance Models for Computer-Aided Engineering*. New York, NY: Academic Press, Inc.

Fotta, M.E., & S. Nicholson (2007). Hemets – Human error modeling for error tolerant systems. *Proceedings of the 14th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, pp 204-209.

Karikawa, D., Takahashi, M., Ishibashi, A., Wakabayashi, T., & Kitamura, M. (2006). Human-machine system simulation for supporting the design and evaluation of reliable aircraft cockpit interface. *SICE-ICASE International Joint Conference*, pp.55-60, October 18-21.

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*. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 105-144.

Lüdtke, A., Osterloh, J-P., Mioch, T., Rister, F., & Looije, R. (2010). Cognitive modelling of pilot errors and error recovery in flight management tasks. In the *Proceedings of* 7th IFIP WG 13.5 Working Conference, HESSD 2009, Brussels, Belgium, September 23-25, 2009, Revised Selected Papers, (pp 54-67).

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.I., & Sarter, N.B. (2003). Towards a model of error management on highly automated glass cockpit aircraft. *Proceedings of the 12th International Symposium on Aviation Psychology* (pp 882-887). Dayton, OH: Wright State University.

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., Young, M.S., Harris, D., Demagalski, J., Marshall, A., Waldmann, T., & Dekker, S. (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*, (pp 567-571).

Stanton, N.A., Salmon, P., Harris, D., Demagalski, J., Marshall, A., Waldmann, T., & Dekker, S. (2003). Predicting pilot error: Assessing the performance of SHERPA. *Proceedings of the HCII Conference*, pp 587-591.

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.

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*.

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*. Boca Raton, FL: CRC Press, Taylor & Francis Group, pp. 213-242.

3.2. Workload and Multi-tasking Models

3.2.1 Introduction

Workload may be defined as the demands imposed by tasks on the pilots' limited information processing resources. Increases in workload can be imposed either as a *single task becomes more difficult*, as hand flying the aircraft through increasingly turbulent weather, or *by adding tasks* to be performed concurrently (multi-tasking). In multi-tasking, when two or more tasks are time shared, those tasks may compete for **common resources** within the information processing system, as when a pilot must read a data link message, and also scan the outside world; or tasks may utilize **separate resources**, as when the pilot is scanning outside (visual), while communicating by radio with air traffic control (auditory input).

Computational models of workload predict the demands on those resources, whether from single or multi-task environments, and are often validated by measures of workload, such as subjective ratings. In contrast, computational models of multi-tasking obviously require the input of multiple (typically two) tasks, and predict the *decrement in performance* of one or both of those tasks, as a function of the task demands (workload of the components) and the extent of their competition for shared resources (Wickens, 2008b). While many models of multi-tasking focus on the decrement of concurrently performed tasks (e.g., how much does communication with ATC degrade concurrent hand flying of the aircraft), others operate under the assumption that the tasks of interest cannot be performed concurrently, so their predictions are of the time at which a task may be performed in sequence. For example to what extent is communicating with ATC postponed by high turbulence, until such turbulence lessens.

In the following, we describe 32 modeling efforts, categorized into five clusters of workload and multi-tasking models, treating both of these areas within this single deep dive because of their

close relationship, the fact that they are sometimes used interchangeably, and the fact that many models are used to predict both workload and multi-tasking fluency.

3.2.2 The MIDAS Model - Channel-Specific Workload

The closest set of studies to truly validate workload models in a NextGen context are those revolving around the MIDAS model developed at NASA Ames Research Center (e.g., Gore, Hooey, Socash et al., 2011, Gore & Corker, 2000a, Gore, 2008). Although this model predicts several variables (generates output) in addition to workload, it contains within it two modules specifically focused on the workload concept. One is the **channel-specific workload** or resource specific workload derived from the multiple resources approach (Wickens, 2008b) where visual, auditory, cognitive and psychomotor (VACP) workload levels are derived and integrated. The set of four channels has more recently been expanded to six, now including both verbal and spatial cognitive channels, and both motor and speech responses; but for convenience below, we will describe this channel approach to workload as "VACP." The second, more recent addition to MIDAS is a **visual attention model (SEEV)**, which will be treated briefly in this deep dive, because this module is closely related to visual workload, the V in the VACP approach. The workload validation research on MIDAS can be loosely divided into to four subsets. The 'Air' MIDAS software (MIDAS v1.0; San Jose State University's version of MIDAS used by Dr. Kevin Corker) utilizes the channel-specific representation to workload with a notion of embedded prioritization, while the most recent version of MIDAS (MIDAS v5; NASA's version of MIDAS) has been applied to the aviation and other domains, implements the option of using either the channel-specific or the SEEV-generated characterization of workload. 'Air' MIDAS is the first incarnation of the MIDAS software while MIDAS v5 is the fifth augmentation of the MIDAS code base.

(1) In one subset of studies (Gore & Corker, 2000a,b), 'Air' MIDAS addressed VACP predictions, associated with a variety of different cockpit configurations, automation implementations and procedures associated with Free Flight (an earlier concept now represented by the concept of self- separation in NextGen). In Gore and Corker (2000a), 'Air' MIDAS predictions of VACP workload profiles were made for 16 different configurations or "experimental conditions" formed by a 2x2x2x2 factorial design. These predictions were analyzed in a conventional ANOVA approach (variance was achieved through the Monte-Carlo properties of the model, with each run producing a slightly different output, analogous to the variance of different subjects in a HITL simulation). Differences (effects of the factors) were examined, and also compared with the output of a different model architecture, the *Integrated Performance Modeling Environment* (IPME). Both models showed a similar pattern of effects; and the effects appeared to plausibly imitate pilot behavior, thus representing a verification that the model performed as expected and that the procedures that were implemented reflected those expected in the Free Flight.

In Gore and Corker (2000b), scenarios similar to those above (Free flight and conventional) were again run with the two models ('Air' MIDAS and IPME), but here the emphasis was on predicted VACP workload for both flight deck and air traffic control personnel. While both models again had equivalent predictions for the flight deck, they differed in their predictions of ATC workload, where the IPME predictions were much higher. An important observation is that

this higher prediction of controller workload by IPME may be a result of the fact that IPME, unlike 'Air' MIDAS, does not contain a **task scheduler** that can move tasks around in time, in order to prevent large workload spikes. This is a characteristic of human performance (Raby & Wickens, 1994; Laudemann & Palmer, 1997, Gore et al., 2013 in preparation; see Section 3.2.4 below) that can lower overall workload, and mitigate particular episodes of peak workload.

In both of these Gore and Corker studies, the model outputs were carefully examined and found to plausibly replicate aspects of pilot performance and workload; a good example of verification. However neither was accompanied by results validation against actual PITL simulation data.

(2) A more recent set of MIDAS studies containing results validation data for NextGen is comprehensively described in the technical report Gore, Hooey, Socash et al. (2011), although various aspects of this effort are also described in several other papers in the published literature (see particularly Gore, Hooey, Haan et al. 2011, for a clear focused description of key results from the validation).

The effort focused on a comparison between an area navigation (RNAV) approach to an airport, and a NextGen approach imposing increased pilot responsibility (and decreased ATC responsibility) for maintaining separation on a closely spaced parallel approach (CSPA). Both of these approach phases were modeled in MIDAS v5, and both were first subjected to what the authors refer to as "input validation". That is, the model task inputs (environmental triggers, flight crew actions, and their sequence), and basic operator primitives, were evaluated by a team of SME pilots to assure that those outputs (of pilot behavior and workload) were plausible, and statistics were provided regarding the agreement between model-derived inputs and SME-generated workload values. This may be described as model verification, captured by quantitative metrics.

The full MIDAS v5 model generated two sorts of workload predictions. Channel specific workload was predicted along the six [visual, auditory, cognitive-spatial, cognitive-verbal, motor and speech; VACCMS] channels. The inputs to the model were derived from the SME estimates (on a 7 point scale) of these channel values associated with each specific cockpit task. MIDAS aggregated (averaged) these values over tasks and time. MIDAS v5 also predicted a form of *qualitative* visual workload, based on scan data and modeled by the SEEV module of MIDAS v5 (Gore, Hooey, Wickens, & Scott-Nash, 2009; See also sections 3.1 and 3.3). It is qualitative in the sense of identifying the source of high scanning demands, whereas the V component of the VACCMS channel vector is quantitative by identifying the total amount of workload across all of the scanned visual channels.

While MIDAS v5 predictions of workload and scanning were generated for both conventional and NextGen scenarios, only the former scenario was validated. For workload, this validation was achieved by identifying a PITL simulation (Hooey & Foyle, 2008) that collected workload data from three pilots on an approach similar to that which was modeled. The three data points of the correlation between predicted and obtained workload were the three sub-phases of the overall airport approach, and the correlation between model predictions of total workload (e.g., averaged over channels) and the mean workload ratings of the pilots was relatively high (r=0.72).

For validation of SEEV, visual scan predictions from the SEEV component of the pilot model were generated for 3 display areas of interest (but averaged over flight phases) and compared with the empirical scan data for these 3 same display areas, harvested from three independently generated separate PITL studies (Mumaw et al., 2001; Huttig, Anders, & Tautz, 1999; Anders, 2001), to generate 9 targets of prediction. The correlation between model predictions and observed scan data was extremely high (r=0.99). However note that unlike the workload prediction validation, the correlation was not over different conditions (phases of flight).

Both workload and scan models within MIDAS v5 were then applied to the NextGen scenario, in which there was increased pilot responsibility for separation (in particular, monitoring wake vortex displays). These were generated by SMEs knowledgeable in NextGen procedures. In both the technical report (Gore, Hooey, Socash et al., 2011) and in Gore, Hooey, Haan et al., (2011), the MIDAS-predicted **change** in both workload and scanning brought about by this NextGen shift in responsibility from ground to air are presented and discussed. However these predictions are not compared against any actual empirical data of pilots flying with the NextGen procedures, so validation is not available.

(3) A related effort to the MIDAS v5 predictions was carried out by Sarno and Wickens (1995) in that this research was funded by the NASA research program that evolved into the workload multi-task model within MIDAS. This validation effort helped to solidify the importance of channel-specific (VACP) workload; however its emphasis was on multi-task performance prediction, rather than workload prediction, and will be discussed in detail in the following Section (3.2.3). We also note here that two other studies to be discussed in Section (3.2.3 carried out in a different program of research (from MIDAS), did validate model-predicted workload against a total (subjective) workload measure, across 8 different display/task conditions. These studies revealed a mean correlation of 0.53 between predicted and obtained workload (Wickens, Larish et al., 1989; Wickens, Harwood et al., 1988), and these efforts will also be discussed in more detail in (3.2.3) below.

(4) See and Vidulich (1998) examined validation of a channel specific VACP workload model. While their target task, a combat aircraft air-ground missile attack scenario, was quite different from the commercial flight deck task, and their subjects were non-pilots, the study is important because of its use of multiple condition predictions. (See Illinois studies in (3.2.3) below). As with the first cluster of NASA-Ames studies that used 'Air' MIDAS (v1; Gore & Corker 2000ab), a set of 8 different conditions were exercised, formed by the 2X2X2 combination of different missile launch display (2 dimensions) and automation configurations. VACP channel values were assigned to different tasks, and then workload was calculated, after these channel values were summed over tasks. These predictions were validated against (a) SWAT subjective workload measures and (b) SART (situation awareness reporting technique); both of these multi-dimensional assessment tools are used heavily within the Air Force laboratories where the study was undertaken. Surprisingly, the results revealed significant correlations between model prediction and the UNDERSTANDING subscale of SART (situation awareness), but not the resource demand subscale. Correspondingly the measures did **not** correlate well with the SWAT subjective ratings. The reasons for this lack of correspondence between model predictions and subjective workload estimates remain unclear.

A corresponding effort was undertaken by Manton and Hughes (1990), to predict VACP workload of helicopter military pilots in a combat mission. Data were collected from video analysis, and VACP time lines created. Task demands within the channels were based on the McCracken and Aldrich scale underlying the VACP approach (Aldrich, Szabo, & Bierbaum, 1989). An assumption was made that demand within a channel greater than 10 created overload conditions (and was assumed therefore to cause performance breakdowns and warrant rescheduling). However no validation was provided. Furthermore, the tasks of combat during helicopter flight (with resultant use of combat displays) are quite different from those imposed on the NextGen commercial pilot.

Finally, Lyall and Cooper (1992) computed (predicted) workload profiles in commercial aircraft pilots by aggregating SME-assigned values in three workload channels (perceptual, cognitive, and psycho-motor), comparing pilot and co-pilot workload on normal versus time compressed departures. While the profiles were presented, no validation data were provided. The prediction was discussed in terms of shared roles and responsibilities between pilot flying and pilot monitoring. This study will also be discussed in the R&R deep dive (3.5).

3.2.3 Multiple Resource Conflict and Multi-Task Interference

Another cluster of studies descend from the multiple resource model developed by Wickens (1980), and its derivative, the WINDEX model computing dual task interference, developed by North and Riley (1989). In the VACP models described in Section 3.2.2, demands within multiple-resource-defined channels were simply summed (or averaged) to reach an overall workload prediction. In the **conflict matrix** approach, demands of two tasks within channels that share more resources in common are more heavily penalized (higher predicted multi task workload and greater dual task interference), than channels sharing fewer resources. Thus for example, two visual tasks will interfere more (higher conflict) than a visual and an auditory task; but the latter pair will still impose a substantial degree of interference because they share demands on common perceptual resources; and if they are both verbal (e.g., reading while listening to speech) this predicted interference will be quite high, but still not maximum (see Wickens, 2002a, 2005, 2008a, and Section 3.2.7) for a more detailed discussion). Furthermore, while the modeling effort described in Section 3.2.2, was explicitly designed to predict "workload" (and hence, validated against subjective workload ratings), those efforts described in the current section are designed to predict **multitask performance interference**, instead of (or in addition to) rated workload. Several studies have explored these techniques, sometimes evaluating and comparing different model computational algorithms for predicting this multi-task interference.

North and Riley (1989) provide a clear demonstration of the calculations invoked in WINDEX, but do not employ an aviation environment.

Miller (1998) employed the WINDEX version of the multiple resource model to make extensive predictions of multi-task flight deck workload. The paper provides an excellent description of the task analysis, and one form of computing conflict interference, while making multi-task workload predictions across a variety of cockpit interfaces. Unfortunately the results of this study were not validated, and the aviation environment was that of a combat helicopter, rather than the

commercial cockpit. A study by Wickens, Bagnall, Gosakan, and Walters (2011) like Miller (1998) provides a description of the application of the MRT conflict matrix to the unmanned air vehicle pilot's task and cockpit design. However this was not validated against PITL simulation data.

A set of three studies in the Illinois Aviation Research Laboratory, all have several features in common, and their general experimental approach is as follows: subjects (usually holding a private pilots' license) performed a low-fidelity flight simulation. In two studies (Wickens, Harwood et al., 1988; Wickens, Larish, & Contorer, 1989) subjects flew a helicopter simulation low over the terrain (a very high workload task), and in the third study (Sarno & Wickens, 1995) they flew a simulated instrument landing system (ILS) landing (a 2-axis localizer and glideslope tracking task). Concurrently subjects in all three studies performed a series of aviation relevant decision tasks that imposed heavy multi-task workload. In the ILS task they were also responsible for a visual monitoring task, designed to increase workload further. The decision tasks typically varied in their complexity, their code (spatial vs. verbal: e.g., vectoring versus fuel calculations) and their display modality (auditory voice versus visual text). In some studies, flight task difficulty was also varied. Thus in each study a wide range of potential interference conditions were created in the 3 or 4 way factorial design (e.g., 8 or 16 conditions) that differed from each other in the extent to which common vs. separate resource demands were imposed between flying and decision making, as well as in the difficulty of the two component tasks themselves. Performance was measured by the most sensitive axis of control, vertical deviations from a target altitude, while single scalar measures of subjective workload were typically assessed.

Then several variants of a multiple resource model predictions of dual task interference were generated. From the simplest to the most complex, these were:

- Simple task time-line model (STL), summing the total time required by tasks and dividing by the time available
- Time-line analysis procedure (TLAP; Parks & Boucek, 1989), in which penalties were added whenever two tasks needed to be performed concurrently. Note that this measure would be 0 if all tasks were performed sequentially, whereas it would not be 0 under the STL algorithm above. The TLAP model discussed more in Section 3.2.3 explicitly penalizes time-sharing.
- Total demand model (TDM) very much like the VACP models described in Section 3.2.1, in which the demand across all channels was summed to derive a total workload.
- Undifferentiated Capacity model (UCM), in which demands were summed, but added penalties were imposed to this sum whenever two tasks needed to be performed concurrently (as with the TLAP model).
- Multiple Resource Model (MRM) in which demand components were added to a *conflict component* that penalized interference to the extent that two tasks demanded overlapping resources defined by the multiple resource model (Wickens, 2008b; e.g., two visual tasks, more than a visual and auditory task).

  The collective results of these studies are summarized in Table 3.2.1 below, which presents the mean correlation (across studies, along with their range to show disparity) between model predicted and obtained measures of either subjective workload or flight task performance decrement. The highest values in each row are highlighted.

Table 3.2.1. Summary of validation efforts against workload and performance decrements.

	STL	TLAP	TDM	UCM	MRM
Workload	No measure	.27 [0-0.55]	**.61[.57-.65]**	.41[.21.68]	.21 [.03-.36]
Performance decrement	0	.50 [.15-.87]	.42 [.08-.74]	.45 [.03-.77]	**.66 [.48-.75]**

What is clear from these data is that models that predict multitask performance data are not the same as those that predict workload (and vice versa). In particular the multiple resource conflict models (MRM) that predict the interference between tasks do well in such predictions, but not well in predicting subjective workload (see also Yeh & Wickens, 1988); whereas models based only on task demand (e.g, summing workload across channels) do a better job of predicting subjective estimates of workload, than of task interference. These conclusions were buttressed by more refined analysis of different model properties carried out by Sarno & Wickens (1995).

Finally, a study by Riley et al. (1991, experiment 1) was quite analogous to the modeling efforts by Wickens, Harwood and their colleagues summarized above. Five models, of increasing levels of sophistication and complexity regarding VACP resource conflict, generated workload predictions of licensed commercial pilots flying a 737 high fidelity simulator. Unfortunately their validation data were restricted to NASA TLX ratings, assigned by SMEs who observed videotapes of the crews' performance. Nevertheless, using these correlations across conditions (N=7 for the captain, N=11 for the first officer), they concluded that the additive conflict version of the multiple resource model was adequate, and accounted for just as much variance as more complex models.

The latest version of MIDAS that incorporates the MRT (Gore, 2013, in preparation; see Section 3.2.1) also now includes both conflict and demand components. This MIDAS version has been verified, but not yet validated.

3.2.4 Time Line Analysis

While time line analysis models were treated briefly above, we also describe them here in more detail because four studies have exclusively examined their predictive capabilities and these studies also represent the origins of the technique. The TLAP technique is best described by Parks and Boucek (1989), who focus on the critical role of the ratio of:

[time required]/[time-available] (or TR/TA)

within pre-defined intervals, to predict workload. When this ratio exceeds 1.0, a period of workload overload may be defined. The authors describe the relationship of these data to a large commercial aircraft full-mission simulation study, carried out jointly between Boeing and Douglas Aircraft companies (Boucek, Sandry-Garza et al., 1987). Importantly, Parks and Boucek report that when the critical ratio of TR/TA exceeds 0.8 (that is, a given interval of time is more

than 80% filled with tasks), "…pilots have been observed to start dropping tasks…"[p. 54] as if this 80% level may form some sort of a "red line" of workload overload in multi-tasking. However the authors offer no specific data source for this observation. Parks & Boucek also compute TR/TA ratios within six separate channels (visual, manual-left, manual-right, verbal, auditory, cognitive), and a weak form of validation is reported between workload in these channels, and accuracy of a secondary task carried out in the simulation (3 significant correlations are reported within the 6 channels), as these were both assessed in the high-fidelity flight simulation (Boucek et al., 1987). We say that the validation was "weak" because the secondary task was not a true flight task; but one meant to assess continuous auditory monitoring requirements (monitoring ATC communications for a call sign).

A second application of Time Line Analysis is offered by Stone, Gulick, and Gabriel (1987) who describe in detail the applications of this technique to a comparison of commercial aircraft flight deck designs. Similar to Parks and Boucek, they break down the time line into specific channels of visual, auditory, left hand, right hand, and feet. There is no cognitive component. Workload is measured by the percent time that each channel is occupied with a task (summed over all tasks) within designated time intervals. Workload within a channel is defined similarly to the method of Parks and Boucek (e.g, TR/TA within each channel). No validation data are offered.

A third time line application is a model developed by Muraoka and Tsuda (2006) designed to predict the sequence of procedures, from a combination of inputs from the flight data recorder, and written standard operating procedures (SOP) provided by the airlines. The model is incorporated in a tool called OPSAMS (Operational Procedures Safety Analysis and Monitoring System), and enables reconstruction of the VACP workload channels. OPSAMS has assumptions built-in regarding rescheduling, should channel-specific workload become excessive (see Section 3.2.6 below; and see also Gore, 2013 in preparation). It is applied to make predictions on three different approach and landing scenarios, routine, time compressed, and one with flaps inappropriately set. No validation data are provided.

The model developed by Laudemann and Palmer (1997) at NASA Ames is also one that essentially depends on a time-based approach (i.e., time-line analysis). Here flight deck tasks are scheduled according to **windows of opportunity** for any given task, and the workload values of any two tasks with overlapping windows are summed. Task urgency is used as a proxy for workload; urgency grows as the end of the window approaches, and this urgency function increases more rapidly for tasks of higher priority. These summed urgency (workload) estimates are integrated over a flight phase (unspecified) to yield a total workload prediction, which was validated in two ways against the data of 36 pilots flying a high-fidelity simulator in a Line-oriented flight training (LOFT) evaluation:

(1) The 6 highest and the 6 lowest performing crews were selected via an algorithm combining flight crew errors with SME evaluations (of videotapes), and the two groups differed significantly on the workload score predicted by the above algorithm.

(2) The overall workload score correlated significantly across crews ($r = 0.42$) with the SME's estimation of crew workload (again, an assessment based on the videotape).

Walden and Rouse (1978) also employed a time-based approach in which they applied formal queuing theory. Unlike the approach of the three time-based models above, all of which enabled concurrent processing of two tasks to be carried out, queuing theory typically models strict single channel (single server) behavior, and the emphasis is on **scheduling** of which task(s) get performed immediately, and which are delayed or deferred. Subjects (non pilots) using a low fidelity desk-top simulator flew the "aircraft" (a simple 2 axis control) and monitored gauges for failures (excessive fluctuation). Event rates and task priorities were used in the model to drive task switching rules, and the model predicted the delay in noticing gauge failures quite well ($r = 0.96$). However the simulation fidelity was quite low, the subjects were not pilots, and validation data were not provided for flight control error. We also note, as described in the previous section, that Sarno and Wickens (1995) provided a pure time-line model prediction of multi-tasking.

Finally, before leaving our discussion of time-based (and particularly queuing) models, we note the close analogy between such models, and predictive models of visual scanning, such as SEEV (Wickens McCarley et al., 2008; Wickens, Goh et al., 2003; Gore, Hooey, Wickens, Scott-Nash et al., 2010**;** Steelman-Allen, McCarley & Wickens, 2011; See 3.2.2 above). These models are time based, in that the eye scans one location at a time; and like queuing models they do not allow concurrent processing (the eye is assumed to look at only one place at a time). Furthermore, although where one looks does not always correspond to the task one is performing, an assumption can be made in highly visual environments like the flight deck, that such a correlation between looking at task related information, and performing that task is reasonably high (Wickens, Bagnall, Gosakan, & Walters, 2011). The SEEV scanning model validation was briefly mentioned in the context of the MIDAS studies in Section 3.2.1 above, the majority of such validation efforts are summarized in the models of situation awareness (See Section 3.3).

3.2.5 Single Task Demand Models

Some models have focused on only one dimension of a task, or one type of task, to provide predictive precision on the workload demands of that single task (even while acknowledging that most of aviation involves multi-task workload). Such models are typically easier to validate. A prototypical example is the model of multi-engine aircraft flight handling qualities, presented and validated by Rickard and Levison (1981) in the context of the Optimal Control Model (OCM), and these have a long history of development and validation in the publications of the *Annual Conference on Manual Control*. Such models will derive the predicted workload of controlling the aircraft from engineering analysis of closed loop flight dynamics, based on such quantifiable characteristics as control order, control lag, feedback loop properties and system response gain (Wickens, 1986). Rickard and Levison note in their validation experiment that "…objective performance measures [flight technical error] and Cooper Harper pilot ratings [a subjective workload scale tailored specifically to handling qualities] were largely consistent with each other ***and with analytic predictions"*** (e.g., model output). Although formal correlations were not presented, they could be derived from the data, and are evident in the graphs presented. A second example of the single task demand model is presented by Parks and Boucek (1989). While these authors emphasize time line analysis (see 3,2,4 above), they also describe a computational model of cognitive complexity based on information theory that is used to predict the workload of flight deck display-control interactions. Validation data are not reported.

Two modeling approaches have directly examined models to predict the complexity of FMS automation, Gil, Kaber et al. (2012, in press) developed a model based on GOMSL (see 3.2.6 below). This particular model, described in more detail in the automation deep dive (3.4), was validated against both physiological, performance (vertical flight path tracking error) and subjective workload measures, in comparing predicted complexity of three FMS designs, against the obtained measures of pilots working with these designs. Significant correlations with both physiological (heart rate) workload measures and performance (vertical deviations) are reported. Predictive correlations were high and significant between some aspects of cognitive complexity and physiological measures, as well as lateral deviations. Specific statistics were not provided for predictions of subjective workload.

Sebok et al. (2012) also developed a model of FMS complexity (also described in detail in the automation deep dive Section 3.4), which was not based on a particular existing architecture, but relied upon a total "count" of certain FMS features known to impose added load (e.g., number of modes, degree of mode interaction). While this model was not directly validated, it contributed to the predicted figure of merit score of the larger FMS model of which it was a part, and this figure of merit did receive some form of validation in the manner described in the automation deep dive Section 3.4.

Eng et al. (2008) developed a Cognitive Constraint Model (CCM), that operates around a CORE (constraint based optimal reasoning engine). While this model is primarily a procedural model, and hence was not given a deep dive, it is also indirectly related to single task workload for the following reason. The authors model the trade-offs that operators make between minimizing the time required by a sequence of cognitive operations, and minimizing the aggregate **working memory demands** of those operations. For example an optimal sequence might involve minimizing the time that large amounts of task related information are placed in working memory by "dumping" this information as soon as possible, rather than holding it while other steps are accomplished. The authors apply the model to two different FMS interfaces; that of a conventional Boeing 777 and the redesigned FDOF (flight deck of the future; see section 3.4.3) observing the predicted reduction in working memory demands of the FDOF. However these predictions were not validated against pilot performance or workload data.

3.2.6 Knowledge based models: ACT-R and GOMS

The ACT-R / GOMS approach is primarily focused on multi-tasking rather than workload, and is primarily a sequential task selection model, rather than a model of concurrent tasking like VACP. As such, models based on these two architectures bear a resemblance to the classic queuing theory models (see Walden & Rouse, Section 3.2.4). However they are considerably more complex in choosing not only what tasks will be done at a given period of time, but also what processing routines within each of two time-shared tasks will be done when the two tasks compete for the same routine that cannot be shared between them (such as a decision, a visual scan or a manual action). (Anderson & LeBiere, 1998; Byrne, Kirlik, & Fleetwood, 2008; Lebiere & Archer, 2008). This approach was described in the Error model deep dive section (3.1). D-OMAR (Deutch & Pew, 2008) also described in that section includes a task selection element for multi-tasking. However its predictions in this regard were not validated.

In implementing this approach, Schoppek and Boehm-Davis, (2004) focus heavily on the knowledge required for different tasks, whether declarative or procedural, and predict different speeds of knowledge retrieval as a function of learning. GOMS (goals operators, methods, selection rules) is a task analytic method that identifies the procedural steps by which tasks are carried out. The two modeling architectures are closely related, and are blended in the ACT-R version of a pilot-automation interaction model called "ACT-FLY" examined by Schoppek and Boehm-Davis (2004). This model, and a similar one developed by Schoelles and Gray (2011) called SimPilot also adopt specific multi-tasking assumptions from the model of **threaded cognition**, developed by Salvucci and Taatgen (2008), a model that describes how different within-task routines are allocated to different tasks (usually two) at any given moment.

Schoelles and Gray describe how their model is applied to a pre-takeoff taxi sequence, and offer neither specific predictions nor validation data. Schoppek and Boehm-Davis provide a more elaborate modeling description (using ACT-FLY") of an automated descent procedure using an FMS. The model outputs an altitude profile and certain procedural errors in using the FMS. Its output was qualitatively validated by comparing these errors against those of six transport pilots, who flew a PS1 desk-top simulator on a descent profile equivalent to the modeled profile. Explicit comparisons of model and pilot output however were not presented in the data to enable quantitative validation.

Polson and Javaux (2001) present a model, discussed in more detail in the pilot-automation interaction deep dive section (3.4), and also heavily related to visual attention models, which predicts why pilots do not often scan the flight mode annunciators, a major issue in FMS monitoring. They apply a GOMS 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. No validation data were provided. Also, as noted above, Gil, Kaufman et al., (2010, 2012, in press) apply GOMS to model FMS complexity.

Finally, although the MIDAS model discussed in (3.2.2) does not contain ACT-R or GOMS architecture, it does contain a sophisticated task management strategy (based on Freed, 2000), described in Gore et al. (2013, in preparation) that is parallel to those described above. This strategy predicts the sequential behavior of scheduling when a workload redline is exceeded. Verification of these model outputs has been carried out, but the task selection and task management features are not yet results-validated against PITL data.

3.2.7 Integrative Summary: Workload and Multitasking

Dimensions of Modeled Processes

Figure 3.2.1 provides an overview of the dimensions of sophistication of workload and multi-tasking modeling efforts.

At the bottom are represented the two sorts of pilot performance entities that are limited, and hence both contribute to workload when they are more demanded, and to a decline in

multitasking when they are excessively demanded: these entities are time, and resources other than time (e.g., mental effort). Moving upward from the base in the figure, time-based models, or the time component of other models (right branch), increase in their complexity according to the sophistication of how time management is modeled (e.g., incorporating task scheduling algorithms; Freed, 2000; Gore, 2013, in preparation). Resource (other than time) based models increase in complexity (left branch) along two sub-dimensions: according to how the demands of individual tasks for resources are modeled (see 3.2.5 above) and predicted, and according to the postulation of multiple resources (Section 3.2.3). In the following, we describe these ordinal scales of complexity and sophistication relating to the models reviewed; but also note that any given model can capture different levels along any or all of the three axes. Thus a particular model may have a simple representation of how time is managed, but a complex representation of resources; and while many complex models of sequential time management tend to make single channel assumptions of task management (one task at a time, no concurrent performance), they need not do so (e.g., Gore, 2013, in preparation; Laudeman & Palmer, 1995).

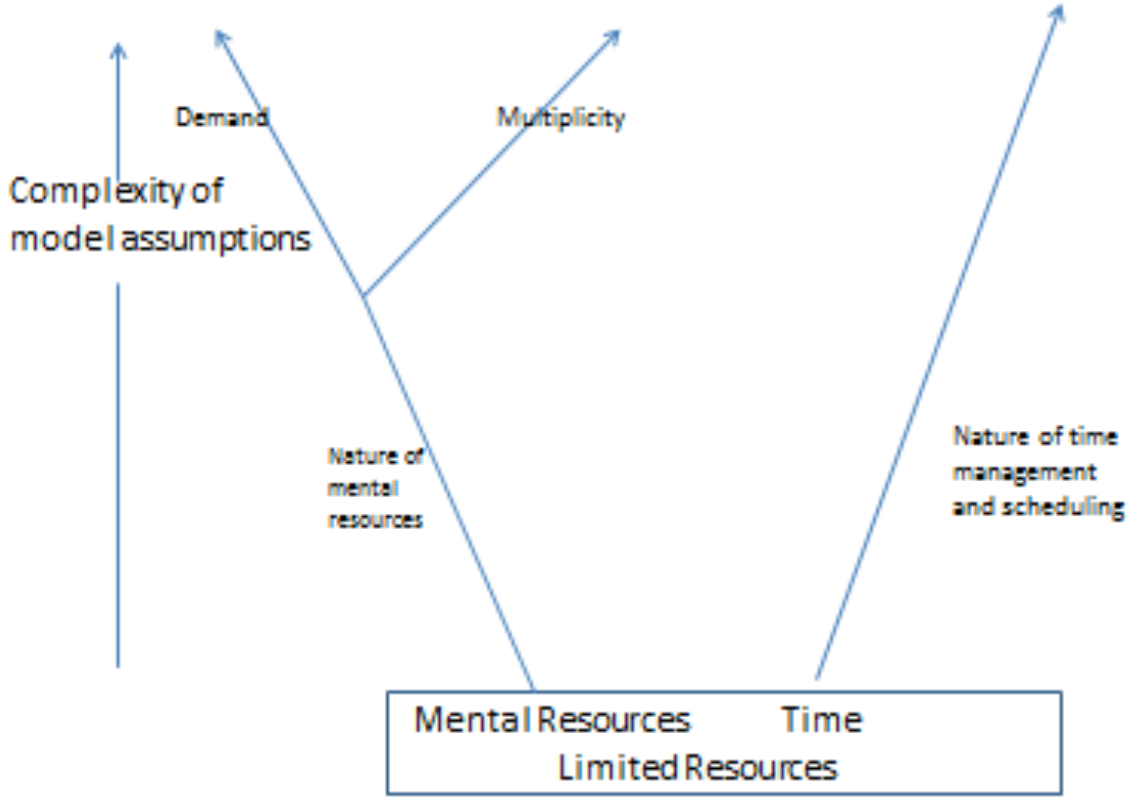

Figure 3.2.1: Complexity levels of workload and multi-task models

The following represents seven levels of increasing model complexity and sophistication.

- **1. Single channel time**. Beginning at the bottom of the figure, models can assume that the only resource demanded is time, and that the time demanded by tasks is well less than the time available, so that concurrence is never required. Such a situation might apply in the middle of a

cruise segment on a highly automated aircraft. We have observed in Section 3.2.3 above that such models do not well predict any situation when concurrent tasking is required (e.g., the STL model in table 3.2.1). However, moving upward along the right branch, such purely time based models can vary greatly in the sophistication and complexity they assume when two tasks do compete for the same time slot. The variation ranges from simple queuing theory (Walden & Rouse, 1978) or scanning models (Wickens, McCarley et al., 2008) to complex scheduling algorithms (Schoppek & Boehm-Davis, 2004; Freed, 2000, Gore, 2013, in prep; Muraoka & Tsuda, 2006).

- **2. Concurrence allowed: all or none**. Moving upward on the left branch of Figure 3.2.1, models assume that tasks can be performed concurrently; but the simplest assumption is that workload is simply defined by *adding tasks*: two concurrent tasks impose twice the workload as a single task. Such models like the TLAP model (Parks & Boucek, 1989) make no assumption about how demanding the tasks are, nor whether they use same or different resources.

- **3. Quantifying Demand.** At this next level of complexity on the left branch of Figure 3.2.1, the resource models are further differentiated. Moving up the left limb, models can vary in the sophistication with which they characterize the demand level of individual tasks (e.g., simple /difficult versus demand level on a 7 point scale). The modeling of flight handling qualities (Rickard & Levison, 1981) is an example as is the proposed urgency/importance concept of Laudeman and Palmer (1995). So too is the measure of task complexity offered by Parks and Boucek (1989) for display complexity, and FMS complexity and by Gil, Kaufmann, et al., (2010) and Sebok et al., (2012) for FMS complexity, as well as working memory load, offered in the CCM of Eng et al.(2008). These demand levels can be predicted for tasks as a whole, or within separate channels, such as VACP, which brings us to the limb on the right side of the left branch: resource multiplicity.

  4. Multiple Resources. Several of the modeling efforts have postulated multiple channels or resources, with the classic VACP channel structure often used by the earlier applications of workload models (Manton & Hughes, 1990; Muraoka & Tsuda, 2005; Parks & Boucek, 1989; Stone et al., 1987; Sarno & Wickens, 1995). We note here that these models are joined with complexity level (3) above, when a task can be associated with a demand level within each of the separate resources populating a given model. Over the last few decades, the 4-channel model of VACP (e.g., See & Vidulich, 1998, Manton & Hughes, 1990) has realized increasing elaboration, to more closely approximate the structure of multiple resources within the brain (Wickens, 2008a; see Gore, 2013 in prep; Gore, Hooey, Socash et al., 2011). When a task is analyzed into its separate resource components, demands can be assigned either 0 or 1, or can be associated with a more graded level of demand, (analogous to the distinction between [2] and [3] above). The source of these channel specific demand levels still remains heavily dependent on SME estimation (e.g., the classic "McCracken & Aldrich scale"); however computational measures of concepts like cognitive complexity (Parks & Boucek, 1989), should allow more objective computations of channel-specific demands to be incorporated in models. Some models of air traffic control have adopted this approach regarding air space complexity (Boag et al., 2006), as have some models of pilot-automation interaction regarding the complexity of the flight management system (Sebok, Wickens et al., 2012; Gil, Kaufmann et al., 2010).

- **5. Providing only the vector of demands.** At a lower level of complexity, the vector of demand values across channels can simply be offered as a workload estimate; and assumptions made that if values within any channel exceed a particular number, this defines "overload" (e.g., summation of demands over tasks within a channel greater than 10; Manton & Hughes, 1990).

- **6. Combining demand values across channels: simple addition**. In contrast to level [5], some models explicitly define workload to be the sum or average across channels (e.g., See & Vidulich, 1998; Gore, Hooey, Socash et al., 2011). If one thinks of workload as being proportional to the total engagement of the pilot's brain, this is a very plausible assumption, but it does not capture the competition between resources inherent in predicting multi-task interference.

- **7. Combining channel values, channel weighting.** Still greater modeling sophistication assumes that not all channel (resource) pairs, populated by a task, will contribute equally to the penalties of multi-task performance (Sarno & Wickens, 1995; Miller, 1998; Gore et al., 2013, in prep.) In particular, two tasks that rely on more similar resources (e.g., two visual tasks, or two verbal language-based tasks) should be penalized more than two that use more separate resources (e.g., a visual and an auditory task). Such weighting can be derived from the dimensional structure of multiple resources within the brain (Wickens, 2008b), whereby two tasks sharing more dimensions, will be weighted more heavily in assessing multi-tasking penalties. There is a diversity of techniques used for producing such differential weighting (Riley et al., 1991).

  At this point it is important to reiterate the distinction between models of workload and models of multi-tasking (performance decrements). Although the term "workload" is often applied to both, they are distinct, both theoretically and practically. The distinction is provided by considering the different model levels in 5, 6 and 7 above. In level [6], demands within channels are added. Hence a pilot may be heavily loaded, with much of her capacity saturated. But because there is no overload in any given channel, and the channels use (somewhat) separate resources; there will be minimal performance decrement. Think of the pilot who is competently engaged in hand flying the aircraft, while carrying on a dialogue with ATC. In contrast, at level [5], only one channel (e.g., visual) may be loaded, as when the pilot must monitor both outside and in. Overall workload may be low (the pilot does not "feel very busy"), but the potential for a dual task decrement is high if, for example, critical visual events occur both inside the cockpit and outside, at the same time. A similar circumstance could arise at level [7]; for example suppose a task analysis reveals a situation in which the pilot must talk and listen simultaneously. Both use verbal resources, and the overload could occur in the verbal resources (potential performance breakdown).

  Recall too that the data themselves shown in table 3.2.1 reveal that the models that best predict workload (e.g., via total demand) are not those that best predict multiple task decrements (e.g., via multiple resource conflict matrix). But it will be noted that the two measures are usually positively correlated, and to some extent affected by the same variables (e.g., number of tasks to be performed concurrently), and hence it is appropriate to consider them both within this single deep-dive section.

Above then, we have arrayed models along an ordinal scale of increasing level of complexity and sophistication. Not surprisingly, more complex models have more complex software, and require greater specialized knowledge to "run". As a partial consequence of this, more complex models tend to be less likely to be fully validated. Facilities that have the expertise and competence of complex software modeling are often less likely to have access to full fidelity flight simulators and pilots. We summarize the state of validation in the following section.

3.2.8 Statistics of Validation of Multi-Task and Workload Models

Thirty-three separate "modeling efforts" of cockpit multi-tasking or workload were identified. These correspond closely to the rows of the spreadsheet. But they involve separate counts if more than one model was addressed within a given paper (e.g., Sarno & Wickens, 1995). Of these, 19 contained models predicting multi-tasking and 15 contained models of workload. Hence, several of the papers modeled (predicted) both commodities. Of the 34 efforts, 11 contain true quantitative validation (typically a correlation between predicted and obtained measure), and three contained a more qualitative evaluation, providing empirical data, and calling attention to the degree of similarity between the predicted and obtained measures. Of the 11 quantitative validations, only two could be said to apply to true NextGen procedures and design (Gore, Hooey, Socash et al., 2011; Gil, Kaber et al., 2009).

When the models in this deep dive are examined from another perspective, ignoring (or independent of) their degree of validation, four of the modeling efforts could be said to apply directly to the NextGen scenario (cockpit and/or procedures), while another seven apply more broadly, in predicting FMS-induced workload issues; (this is assuming that the FMS, or similar flight path automation tools will be heavily involved in NextGen operations. In this regard, the reader should also consult the Automation deep dive for FMS models that do not directly predict workload). Finally, we also note that of the 29, seven address either helicopter piloting, or military combat aircraft missions, where models may generalized to the NextGen commercial cockpit, but specifics of the application do not.

In closing, we remind that the summary statistics above do not include the models of visual attention, such as SEEV, which could be considered to predict sequential multitasking in environments such that where the pilot looks corresponds to the tasks that the pilot performs. These models have a slightly better record of validation within the aircraft cockpit, but are not summarized in depth in this report (see 3.2.1).

3.2.9 References for Workload and Multi-Tasking

Boag, C., Neal, A., Loft, S., & Halford, G. S. (2006). An analysis of relational complexity in an air traffic control conflict detection task. *Ergonomics*, *49*(14), 1508-1526.

Deutsch, S.E., & Pew, R.W. (2008). D-OMAR: An architecture for modeling multitask behaviors. In D.C. Foyle & B.L. Hooey (Eds.) Chapter 8: Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group, pp. 183-212.

Eng, K., Lewis, R., Tollinger, I, Chu, A., Howes, A. & Vera, A. (2008) Generating automated predictions of behavior strategically adapted to specific performance objectives. *CHI 2008 Proceedings, Automatic Generation and Usability* Montreal Can.: Association for Computing Machinery.

Gil, G., D. Kaber, S. Kim, K. Kaufmann, T. Veil, & P. Picciano (2009). Modeling pilot cognitive behavior for predicting performance and workload effects of cockpit automation. *Proceedings of the 2009 International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, 124-129.

Gore, B.F. (2008). Human performance: Evaluating the cognitive aspects. *Handbook of Digital Human Modeling* (Ch. 32, pp. 1-18), NJ: Taylor and Francis.

Gore, B.F. & Corker, K.M., (2000a). Human performance modeling: Identification of critical variables for national airspace Safety. In the Human Factors and Ergonomics Society Annual Meeting Proceedings. Santa Monica, CA: HFES.

Gore, B.F. & Corker, K.M., (2000b). Value of human performance cognitive predictions: a free flight integration application. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: HFES.

Gore, B. F., Hooey, B. L., Haan, N., Bakowski, D. L., & Mahlsted, E. (2011, July 9 - July 14). A methodical approach for developing valid human performance models of flight deck operations. *Paper presented at the Human Computer Interaction International* (HCII) 2011, Orlando, FL.

Gore, B. F., Hooey, B. L., Socash, C., Haan, N., Mahlsted, E., Bakowski, D. L., Gacy, A.M., Wickens, C.D., Gosakan, M., Foyle, D. C. (2011). *Evaluating NextGen closely spaced parallel operations concepts with human performance models*. HCSL Technical Report (HCSL-11-01). Moffett Field, CA: NASA Ames Research Center.

Gore, B. F., Hooey, B. L., Wickens, C.D., Socash, C., Gacy, A.M., Brehon, M, Gosakan, M., Foyle, D. C. (2013, in process). *The MIDAS workload model*. Internal report. Moffett Field, CA: NASA Ames Research Center.

Laudeman, I.V., & Palmer, E.A. (1995). Quantitative measurement of observed workload in the analysis of aircrew performance. *The International Journal of Aviation Psychology, 5(*2), 187-197.

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*. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 105-144.

Lyall, E. A., & Cooper, B. (1992). The impact of trends in complexity in the cockpit on flying skills and aircraft operation. In the *36 th Human Factors Society Annual Meeting, Atlanta, GA* (pp. 1181-1184).

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

Miller, C. A., (1998). *Case Studies Involving W/Index*, Honeywell Technology Center.

Muraoka, K. & Tsuda, H. (2006). Flight Crew Task Reconstruction for Flight Data Analysis Program. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50(*11): 1194-1198.

Parks, D. & Boucek, G. Workload prediction, diagnosis and continuing challenges. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R. Sutton, & L.V. Breda (1989). *Applications of*

human performance models to system design (Defense research series, Vol. 2). New York City, NY: Plenum Press.

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

Rickard, W. W., & Levison, W. H. (1981). Further tests of a model-based scheme for predicting pilot opinion ratings for large commercial transports. *Proceedings of the 17th Annual Conference on Manual Control,* pp. 247-256.

Riley, V., Lyall, E., Cooper, B., & Wiener, E. (1991) *Analytic methods for flight-deck automation design and evaluation. Phase 1 report: flight crew workload prediction*. FAA Contract DTFA01-91-C-00039. Minneapolis Minn: Honeywell Technical Center.

Sarno, K., & Wickens, C. (1995). The role of multiple resources in predicting time-sharing efficiency: An evaluation of three workload models in a multiple task setting. *International Journal of Aviation Psychology, 5*(1), 107-130.

Schoelles. M.J., & Gray, W.D. (2011). Cognitive modeling as a tool for improving runway safety. The *Proceedings of the 16th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, 541-546.

Schoppek, W., & Boehm-Davis, D. A. (2004). Opportunities and challenges of modeling user behavior in complex real world tasks. *MMI interaktiv*, *7*, 47-60.

Schurr, N. (2011). ALARMS: Alerting and reasoning management system. *Presentation delivered to the 2011 NASA Aviation Safety Technical Meetin*g, St. Louis, MO.

Sebok, A., Wickens, C., Sarter, N., Quesada, S., Socash, C., Anthony, B. (2012). The Automation design advisor tool (ADAT): Development and validation of a model-based tool to support flight deck automation design for nextgen operations. *Human Factors and Ergonomics in Manufacturing and Service Industries, 22*(5), 378-394.

See, J.E., & Vidulich, M.A. (1998). Computer modeling of operator mental workload and situational awareness in simulated air-to-ground combat: An assessment of predictive validity. *The International Journal of Aviation Psychology, 8*(4), 351-375.

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

Stone, G., Culick, R. & Gabriel, R. (1987) Use of task timeline analysis to assess crew workload. In A. Roscoe (Ed) *The practical assessment of pilot workload*. NATO AGARDograph #282.

Walden, R.S., & Rouse, W.B. (1978). A queueing model of pilot decision making in a multitask flight management situation. *IEEE Transactions on Systems, Man and Cybernetics*. pp.867-875, December 1978.

Wickens, C.D., Bagnall, T., Gosakan, M., & Walters, B. (2011). A Cognitive Model of the Control of Unmanned Aerial Vehicles. The *Proceedings of the 16th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University. 535-540.

Wickens, C.D., Harwood, K., Segal, L., Tkalcevic, I., & Sherman, B. (1988). TASKILLAN: A simulation to predict the validity of multiple resource models of aviation workload.

Proceedings of the 32nd Meeting of the Human Factors Society (pp. 168-172). Santa Monica, CA: Human Factors Society.

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

Wickens, C.D., Larish, I. & Contoror, A. (1989). Predictive Performance Models and Multiple Task Performance. Proceedings of the Human Factors Society 33rd Annual Meeting, pp 96-100.

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. Chapter 9 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 213-242.

3.3 Situation Awareness Models

3.3.1 Introduction

Situation awareness in the cockpit for NextGen collectively represents the pilot's understanding of the dynamic current and future state of the aircraft, including the state of automation systems, the state of engine systems, the progress of 4D navigation along a planned route, and the state of its trajectory relative to hazards; weather, traffic and terrain. This may also include the awareness or understanding of the knowledge and workload of other human agents as well: what is ATC's assumption about "who is in charge?" what is the pilot's understanding of the co-pilot's current role and responsibility? And what is the pilot's understanding about the navigational intentions of close proximity traffic?

While situation awareness concepts and measurements of many of these dynamic aviation constructs have been relatively well studied; and some predictive models of SA have been developed in other non-aviation contexts (e.g., Banbury & Tremblay, 2004), the intersection of predictive models of SA and the flight deck context is sparsely populated, and the population is diminished still further in the context of NextGen operations.

The most commonly accepted definition of SA comes from Endsley (1995) and is expressed in three levels representing a progression of information understanding and need. Level 1 SA is perception of elements within the environment. Level 2 SA is the comprehension of their meaning. Level 3 SA is the projection of their status in the near future. In the context of NextGen cockpit traffic scenarios, the three SA level might be expressed as:

- Detection – Aware that there is traffic
- Integration – Aware of the relative location and trajectory of traffic
- Anticipation – Aware of the future location and trajectory of traffic

From the standpoint of the Error modeling deep dive (3.1) the three levels express:

- Level 1 SA: Failure to correctly perceive information (e.g., data are not available, data are difficult to discriminate or detect)
- Level 2 SA: Failure to correctly integrate or comprehend information (e.g., poor mental model, over-reliance on default values)
- Level 3 SA: Failure to project future action or state of the system (e.g. over-projection of current trends, failure to consider all data)

In this section, we describe five clusters of SA models. The clusters are based on commonalities or trends across the modeling efforts. Most papers are included in only one cluster but a few papers include aspects discussed in more than one of the topic areas.

3.3.2 ACT-R / Cognitive Modeling of SA

ACT-R (Atomic Components of Thought – Rational) is a cognitive modeling architecture that has been used to represent a range of aspects associated with human performance in NextGen. ACT-R represents performance at the level of cognitive activity every few hundred milliseconds. It includes asynchronous modules that represent perception, attention, long-term memory, goal states and condition-action rules (i.e., if condition – then action). Due to this level of cognitive decomposition, ACT-R is often combined with a higher-level task modeling environment. The high-level tool is commonly used to represent process steps in a cockpit environment that are informed by the cognitive model. The use of ACT-R represents an attempt to model pilot SA from the bottom-up based on the cognitive functions of the brain.

The paper by Boehm-Davis et al. (2002) describes a use of ACT-R combined with an NGOMSL (Natural Language GOMS) task-based pilot model executing a descent sequence on a desktop flight simulator. This combination of tools was discussed through several papers in the Error deep dive section (3.1) of this report. This paper does not describe the specific characteristics of SA represented by the ACT-R model. Rather, two observations made during model runs are suggestive of SA related errors. The first is the lack of awareness of uncommanded mode changes in the flight management system. The second is an occasional inability to recall the previous goal state following an interruption. The lack of awareness of a mode change is a perceptual failure (SA Level 1) and the inability to recall the goal state represents a general SA failure. No specific performance measures were provided for a validation. However, interventions added to the simulated cockpit interface designed to address the two SA-related observations seemed to result in fewer errors when commercial pilots flew the same descent sequence with the simulator. This would imply some level of verification that the predictions made by the model at least represent possible problems for pilots.

The model presented in the paper by Keller et al. (2004) combined Improved Performance Research Integration Tool (IMPRINT) and ACT-R to represent the pilot's understanding of the aircraft's position relative to adverse weather (SA Level 2). The IMPRINT model executed time-based flight actions associated with using a cockpit weather information system. Audio and visual inputs were encoded as chunks of spatial weather information including bearing, distance, landmarks and relative position. These chunks were 'sent' to the ACT-R model from the tasks in the IMPRINT model that represented the pilot's access to that information. As such, Level 1 SA was assumed to have occurred but was not specifically modeled. The ACT-R model represented

the effects of memory decay, cue priming and rehearsal on the activation levels of these weather knowledge chunks over time. The pilot's weather SA was then based on the activation levels of chunks of weather information at any given time during the simulation. As part of a separate effort, SA data had been collected from pilots using the cockpit weather system that might have been used to compare with model outputs including latency response, recall probability and the magnitude and distribution of positional errors. However, due to various issues the data were not available during the study and no validation of the model results was possible.

3.3.3 Bayesian Networks / Petri Nets / Linear Regression

Five papers within this category present three different computational models of SA. The models are built to represent a particular set of data through various weighting schemes and structures. In each case, the models integrate information elements from different sources into higher levels of SA. The weightings from Bayesian and Petri networks are modeler-assigned usually based on expert knowledge. The concept element relationships from the linear regression model are derived from experimental data.

Two papers (McNally 2005, Zacharias et al., 1996) present the Situation Awareness Model for Pilot In-the-Loop Evaluation (SAMPLE) as a computation model that uses Bayesian Networks to model components of SA. In McNally (2005), SAMPLE is used in conjunction with an aircraft dynamics model (EAAGLES). Sensor data from the aircraft model is used by SAMPLE to assess the current situation and make decisions about pilot actions that are then supplied back to the aircraft model. As such, the system is described as an SA-based decision model. SAMPLE uses an information-processing step to transform sensor data into semantic variables representing events from the simulation (SA Level 1). A situation assessment step then integrates and interprets these variables, using Bayesian networks, resulting in an assessment of the situation (SA Level 2). The events are represented as the lowest level nodes of the networks. Activation of these nodes propagates up through progressively higher associated situation nodes that combine to represent the situation. The event/situation node associations and probabilistic relationships within the network are developed based on expert knowledge of the represented scenarios. Rule-based decision-making is used to then form actions based on the situation (SA Level 3). SAMPLE has primarily been used with military air-to-air combat scenarios. While outputs from the system include a number of SA and performance metrics that should be suitable for comparison with HITL trials, no validation studies were found.

Two papers (Stroeve et al., 2009; 2011) presented the use of the TOPAZ model and its use in assessing risks of runway incursions. These papers were presented in the Error deep dive section (3.1). The modeling of SA is very general and represented as periodic visual inspections of the runway area by the pilots of two aircraft and one air traffic controller on a stochastic time interval. From the pilot standpoint, these inspections represent a portion of the scan path that includes the area outside the aircraft. If the visual scan coincided with the presence of an aircraft in a position to interfere, it was assumed that the pilot or air traffic controller would gain that SA. TOPAZ functions as a multi-agent dynamic risk model based on a hierarchically structured Petri net formalism. Across the Monte Carlo runs of the model, the frequency of these inspections affects the overall probability that the incursion is noticed in time to prevent a collision. As mentioned in the Error section, the reports include only verification rather than validation. Part

of the difficulty is the lack of available data for such a low-frequency event as runway incursions.

A paper by Svensson and Wilson (2002), presents a SA model based on linear regression of data collected during actual flights of military fighter intercept scenarios (i.e., empirically assigned weights). The data were collected from pilots through questionnaires on mission complexity, mental workload, mental capacity, SA, and performance. The analysis revealed statistically significant relationships between seven indices representing the pilot responses including mission difficulty, information complexity on two different types of cockpit displays, workload, cognitive capacity, SA, and performance. Mission complexity was only described as a range from simple training scenarios to applied missions of 'very high complexity' and information load as displayed on both the tactical situation and target indicator displays. The result is a causal flow model of relationships between these elements showing that mission complexity affects workload and that workload, in turn, affects SA and performance. The authors report a strong connection between the information load on the displays and workload and that increases in workload and display complexity both decrease SA. The data from subsequent PITL simulated flights was used to generate another causal flow model and the two models show the same types of relationships between elements. The report presents this as validation but does not provide any numerical comparison of the relative weightings between the model elements. The use of simulated fighter intercept scenarios, which, while not directly applicable to NextGen, at least included data gathered for pilots in both actual and simulated flight environments. In addition, the mission complexity factors in the scenario were, at least partially, expressed in the information load of cockpit displays.

3.3.4 SA Results from Workload Model

The paper by See and Vidulich (1998), describes a Micro Saint model designed to predict pilot workload rather than SA (See 3.2.2). However, the paper includes correlations with pilot-in-the-loop studies of workload and SA and workload predictions from the model. Since the model is primarily focused on workload, it is covered in detail in that section of this report. The workload model covers the tasks associated with aviation combat missions. Twelve subjects (non-pilots) flew the same missions in a flight simulator and both SART (an SA scale) and SWAT (a workload scale) rating were collected. The two main model predictions of overall and peak workload were significantly correlated with the SART SA ratings. Higher SA ratings from SART were associated with model predictions of lower workload. Also, the strongest correlations with the three dimensions of the SART scale were with the Understanding dimension, which is the most representative of SA. This result is illustrative of the strong connection between workload and SA (Wickens, 2002b, 2008b). While the model presented in this paper is not a model of SA, it is a computational model predicting results that are strongly correlated with SA within a military cockpit environment.

3.3.5 Actual Situation versus Pilot SA

Several papers describe efforts at modeling SA by explicitly representing an actual situation and the pilot's understood situation. This SA modeling concept is generally based on the definition that SA is the gap or difference between the real world state and our understanding of that state.

The paper by Donnelly et al. (1997) proposes the Integrated Decision Model (IDM) representing SA, decision-making, and errors based on the Recognition-Primed Decision (RPD) model (Klein et al., 1993). The proposed model uses the comparison of the actual situation and the pilot's mental representation to represent the decision sequence. Three different decision paths used by pilots are included; (1) determining that more information is needed, (2) forming an intention to act with assessment of potential consequences and, (3) when the task is routine or under time pressure, reacting automatically without assessment of choices. Level 1 and 2 SA are represented by the information and interpretation of the actual situation. Level 3 is represented by the action intentions. The paper provides only a proposal for the model without any information about actual pilot scenario simulations or validation efforts and does not appear to contain a computational element.

The paper by Karikawa et al. (2006) presents a model, Pilot Cognitive Simulation (PCS) that models full pilot cognitive capabilities for pilot flying (PF) and pilot not flying (PNF). The Error components of PCS were discussed in the Error deep dive section of this report (3.1). The model is based on a process sequence between the interface and the cognitive functions of the pilots. Pilots respond to required actions based on attention channels defined by VACP and limited by tasks already being executed. Pilot actions are based on agent groups that represent stored knowledge elements, procedural knowledge, information acquisition through monitoring, communications, visual and auditory perception and operational control. PF and PNF share responsibility for some tasks and SA is acquired from each task executed by either PF or PNF. This represents a concept of shared SA as neither pilot has the whole picture. The SA component is based on an internal situation model (ISM) of each pilot and an actual situation model (ASM). SA is represented within the ISM based on inputs from the ASM as information acquired from the cockpit environment (Level 1 & 2) and predictions and actions performed by the pilot (Level 3). Erroneous actions are presented as discrepancies between the SA of the ISM and the ASM. The PCS model was exercised using descent scenarios in which the aircraft was cleared to a new altitude that may or may not be appropriate given nearby terrain. As discussed in the Error section (3.1), it was unclear how model predictions for errors were compared with actual data. While no specific description of validation of the SA components was described, it was reported that PCS was able to detect a plausible problem with a simulated enhanced cockpit display, thus reflecting a sort of verification process.

Two papers, Shively et al. (1997) and Burdick and Shively (2000) describe the original SA modeling components of MIDAS (in MIDAS v2). More recently, aspects of the A/SA model described in the following section has been added to augment the SA modules of MIDAS v5 (Hooey et al., 2010; discussed in Section 3.3.6 below) but the structure of the SA components and how they interact with the other HPM components within MIDAS has not changed. In MIDAS Situational Elements (SE) are the components of the environment that define the situation. Each SE is assigned to context-sensitive nodes that are collections of semantically related SEs. In Shively et al. (1997) the modeled scenario involved an attack helicopter. SEs such as radar, altimeter and airspeed were grouped into an 'own-ship' node. Acquisition of SE components is governed by perception and attention modules but can also be a function of experience or previous knowledge. An SA management function uses the information about each SE to function as a behavioral regulator. For any given point in a scenario, the SA manager

determines which SEs have been acquired at differing levels and compares that against the total number of obtainable SEs. It also determines if any SEs have been erroneously acquired or perceived such that the pilot 'thinks' he knows something but is incorrect. This SA error can be considered a representation of overconfidence in SA (Sulistyawati, Wickens & Chui, 2011). Actual SA is then calculated as the perceived SA minus the error SA versus the total obtainable SA at the time. Each SE and each context node are weighted to express its relative importance to the overall situation. The SA manager uses these weightings to determine the level of SA achieved and to derive subsequent decision-action behaviors within MIDAS.

In Burdick and Shively (2000), two different validation efforts for the computational model of SA (CSA) within MIDAS are presented. (1) The first study compared MIDAS CSA predictions versus ten GA pilots executing six pairs of trials in a low fidelity simulator. Each trial pair differed only in the availability of various SEs making up three higher order nodes for navigation, system management and flight control for a 'high' SA condition and a 'low' SA condition. SAGAT and SART were used to compare the pilot trials with model predictions. The paper presented only summary results showing basic predictive adherence between model and experiment across various SA measures. Two, in task map related questions and unprompted waypoints were significant ($p<.05$). The results showed that CSA could predict both the subjective as well as performance measures of SA but only to a low-resolution level of 'low' or 'high' SA. (2) The goal of the second validation study was to extend the validation to three levels of predicted SA (low, medium, and high). Six GA pilots flew medical evacuation transport scenarios in a rotorcraft part-task simulator. The scenarios included navigation to patient and hospital locations and SA was manipulated by varying the level of route and destination information provided by the flight management systems. Both SAGAT and SART results matched the trend direction for model predictions across three levels of SA, but this was only a qualitative validation. The authors point out that all of these validation results were preliminary.

3.3.6 SA and Visual Scanning

The Attention-Situation Awareness (A/SA) model, developed by Wickens, McCarley et al., (2008) is an extension of the work at NASA-Ames carried out by Shively and his collaborators described in the previous section. Functionally, it is a model of level 2 understanding of the status of dynamic systems, both those related to aircraft stability (aviating) and to 4D navigation (toward desired waypoints, and away from traffic, terrain and weather hazards).

A/SA contains two major components, a visual scan component, and a memory for state component. The scan component is the SEEV model of visual attention (Wickens, Goh et al., 2003; Wickens, McCarley et al., 2008), described briefly in the workload deep dive (3.2.2). This predicts the scan path around the dynamic instruments and outside world that are relevant to updating SA regarding the status of those "situational domains". In the SEEV model these are referred to as Areas of Interest (AOI). As they pertain to situation awareness, they are referred to as "situational elements" (SE) (Hooey, Gore, Wickens, Scott Nash et al., 2011). A glance to a SE (for example a traffic display) will update traffic (situation) awareness to various levels, depending on how long the eye dwells there, with long dwells typically leading to maximum SA. (However poor or ambiguous displays may lower this value from maximum, imposing what the

model describes as data limits.) SEEV has been validated against the scan pattern of users in various environments including the pilot's scan of instruments in the cockpit (Wickens et al., 2003; Wickens, McCarley et al., 2008; and particularly the automated cockpit (Steelman-Allen, McCarley & Wickens 2011, validation 1; Gore, Hooey, Socash et al., 2011). Thus the validation of scan path predictions may be considered as a partial validation of level 1 SA; given that fixation is necessary (although not sufficient) for noticing.

The memory component then initiates a decay of the quality of situation awareness the instant the eye leaves the SE in question, a decay following the properties of long-term working memory (Ericssen & Kintsch, 1995; Durso, Rawson & Girotto, 2007), diminishing to near zero after about a minute; but reset to maximum (minus possible data limit) upon return of the scan to either that SE, or another SE that supports the same situation. For example, scans to both the cockpit window in VMC and to the altimeter, can support altitude awareness on a visual approach.

The A/SA model has been applied to three flight deck situations. In Sebok et al. (2006) it was applied to predict the pilot's awareness of wake vortices on a parallel landing approach, using different display formats. No validation was carried out, but the model outputs were verified by professional pilots. In Wickens, McCarley et al. (2008) the model was used to predict navigational awareness on the runway surface. This was part of the error modeling project at NASA Ames (Hooey & Foyle, 2008) and was described in some detail in the error deep dive. Predictions were made that the loss of awareness of the location on the runway surface at Chicago O'Hare airport, would lead to incorrect turns, and indeed these predictions were born out by a qualitative validation, in which predictions were compared with actual taxiway errors produced by professional pilots taxiing in a high fidelity simulation. Both the model and the obtained error data revealed fewer errors with advanced surface navigation displays.

While both of the above model simulations were of a stand-alone version of A/SA, a third simulation (Hooey et al., 2010) was carried out when A/SA was embedded in the MIDAS v5, described in more detail in Section 4 (Gore et al., 2013, in preparation). The simulation involved a two pilot crew on an approach with either of two cockpits, a conventional cockpit, and a NextGen cockpit in which the captain was equipped with a HITS displayed on a HUD, while the first officer was equipped with an advanced navigational display. The verification of A/SA revealed that the model was sensitive to differences in display configuration and pilot responsibilities. Validation of SEEV-predicted scan paths was accomplished by comparing data with three other scan studies (see 3.2.2). Validation of the full SA model predictions has not been carried out.

3.3.7 Summary and Conclusions

Levels of SA represented in the models

SA models seem to come in two forms. Some treat SA as a single representation of awareness at a specific time either as a general estimate of overall pilot awareness or associated with specific settings or measures (e.g. current mode settings, relative aircraft position). Other models have addressed one or more of the SA levels from Endsley's definition often providing a combination of visual perception model (Level 1) and attention (understanding) model (Level 2) as

representations of information acquisition. Throughout the deep dive reviews of the SA modeling efforts, we have made an attempt to categorize the models by these two different definitions of SA. In many cases, the authors specifically used these definitions and provided their own categorization. However, in many cases we were forced to apply an interpretation based on the modeling details presented. In the case of the three categories associated with the Endsley definition, it was usually clear enough even if the 'Levels' were not specifically stated. For example, any SA model that included a 'perceptual' component was generally categorized as representing Level 1 SA.

General SA

As indicated, many of the model efforts didn't attempt to go beyond treating SA as a single entity of general pilot awareness. A number of characteristics are associated with models included in this category. The predicted value of general SA is sometimes represented as a relative value (e.g. low, medium, high) that can change over time depending on the scenario represented. In some cases, the modeling environments included only a small component of SA as part of a larger cockpit modeling environment. One model didn't attempt to model SA specifically but used a proxy aspect such as workload to predict SA. To the extent that the SA predictions from these general models have been validated, they should be useful as measures of the direction of change of SA for new system designs. At a high level, it will be very useful to know if NextGen design concepts are resulting in a positive (or negative) change in SA even when a more detailed assessment of specific SA components is not supported.

Level 1 SA – Perception

The perceptual component of SA is the component most often represented within the cockpit models. Level 1 tends to be represented in two different ways. One common method is as a general 'perception' process by which information elements are acquired from the environment and included within the SA components of the model. The other method is to use a specific model that emulates the actions of perception, usually visual. In the cockpit, this included various levels of replicating the pilot's scan pattern across instruments and out the window and the prediction of noticing (perceiving) visual events (Wickens, McCarley et al., 2008; Wickens, Hooey et al., 2009). This has the advantage that perception can be objectively measured for validation by scan path analysis. In both cases, when a distinction is made, the information elements are generally those that can be acquired visually. However, several models extend perception to include auditory and knowledge-based elements. In nearly every model, the representation of Level 1 SA is then tied to some level of situation comprehension such that Level 2 SA is also included.

Level 2 SA – Comprehension

Nearly every model that included a concept for Level 1 SA tied that closely with a representation of Level 2 SA. The representation of Level 2 generally occurred in two ways. The first is the use of proxies such as visual fixation or dwell times on a particular cockpit instrument to infer the understanding of the attended information element. While not a perfect indicator of understanding, this has the advantage of using simple time components within the modeling environment. This assumption is also commonly used within eye-tracking studies. Some models stop there and state that the level of understanding for the attended element is static while other models have gone one step further and modeled a time-based change in the quality of the

understanding of an information element (Hooey et al., 2010). During the initial acquisition of any given element the SA level is assumed to be high; and higher with longer fixations. The quality of that information item or ability to mentally access it then degraded over time using various decay algorithms. Many models also then represent an increase in the SA for an information item when next it is attended to within the scenario sequence.

Level 3 SA – Projection

Level 3 SA is the stage that is the least well represented by current models. Those that make an attempt usually focus on generating decisions from rule-based algorithms for a specific situation having used predictions of Level 1 and Level 2 SA to generate the 'situation'. The jump from situational understanding to decision represents a somewhat valid decision-making step similar to Klein's (1989) concept of recognition-primed decision-making. However, from a modeling standpoint, resulting decisions or actions are simply another use of a proxy that does not represent the mental process of combining an understanding of the situation with experiential knowledge to create a **projection** of a future state. A study by Sulistyawati et al. (2011) used SAGAT to assess all three levels of SA of fighter pilots for an air combat simulation. In addition to the SA levels, the pilot's confidence in their SA was also collected. The results show that pilots had the lowest accuracy and highest overconfidence associated with predictions (Level 3 SA); but a model predicting this result was not presented. This result emphasizes the need to focus on supporting the pilots with systems designed to enhance their ability to project future states and support system designers with models that can effectively represent and predict Level 3 SA.

Modeling Distributed SA

The commercial cockpit is fundamentally a two-person shared-task environment in which the crew needs to manage their physical and mental resources. As such, the overall understanding of any situation is a function of the combined SA of the two pilots. The majority of computational SA cockpit models reviewed here are focused on predicting the SA of a single pilot. Only two of the models have included the concept of differing or distributed SA across the pilot flying (PF) and pilot monitoring (PM; Karikawa et al., 2006, Hooey et al., 2010). Both models represented the concept that the two pilots share many of the tasks and while most of the task assignments are predetermined, it is sometimes the case that one pilot will 'back-up' the other if the workload requires. In the models, various actions or tasks occur over time and get assigned to the currently available pilot. The resulting SA components associated with those tasks are then acquired by the pilot who performed the task. In this way, the available SA is split between the crew and the model can express any resulting issues when subsequent tasks executed by one pilot require the SA currently held by the other. One of the models (Hooey et al., 2010) used a specifically NextGen-like topic when the PF was modeled using an updated set of flight instruments (HUD) while the PM was modeled using current day systems. The model, discussed also in the roles & responsibility deep dive (Section 3.5) was able to express not only the components of distributed SA but also the difference in the quality of the SA across the different cockpit systems.

NextGen Concepts within Modeled Scenarios

The SA models described here cover flight topics across a range of both military, GA and commercial aviation scenarios. The papers were specifically limited to those that presented a

computational model of the pilot in the cockpit. However, not all of them relate specifically (or even generally) to NextGen concepts. The SA needs of the fighter pilot in tactical situation represent considerable interest to military system designers. Military-focused SA models include intercept, air-to-air, and air-to-ground scenarios. This need, along with the associated available funding for military research and development, are likely the reasons why so many of the models of SA are either focused on or exercised using fighter-based scenarios. While the SA methodologies presented in these efforts may be valid and useful, the models have not necessarily been applied to NextGen concepts.

While the military-backed SA models are common, models focused on commercial aviation and NextGen-like concepts are not lacking. Two models focused on scenarios related to issues of navigation during flight. One was focused on cockpit systems supporting waypoint navigation and the other was focused on awareness of aircraft location relative to adverse weather. Two modeled SA issues for runway incursions and taxiway turn errors. The most commonly modeled scenarios focused on the approach phase of flight including the descent sequence. Within this context, two models also focused on the very specific NextGen concepts of advanced cockpit systems that included wake vortex information and a synthetic vision system presenting a highway-in-the-sky. This concentration of approach phase and runway issues represents the understanding of the high workload environment and high consequences associated with the concentration and reduced spacing of aircraft at or approaching an airport.

3.3.8 Statistics of Validation of SA Models

Fifteen separate modeling efforts of cockpit SA were identified. These are shown in Table 3.3.1. Of these 15, only one contained a quantitative validation that included a correlation between a predicted and an obtained measure. Two included more qualitative validations in which they pointed out similarities in trends or values between predictions and obtained measures. Of these three efforts, only one can be said to apply to NextGen concepts. Six of the modeling efforts that did not include any validation did report some level of verification. These ‘verifications’ took the form of checks by SMEs on the ability of the model to replicate or output results that ‘appear’ correct or that match a simulated environment in some way. All six of these efforts apply to either general or specific NextGen concepts. Independent of verification or validation, eight of the SA modeling efforts included some application to NextGen concepts while the other three are focused on military applications.

Table 3.3.1 Validation status of SA models.

Model	Verification / Validation	Scenario (s)	SA Prediction
ACT-R / NGOMSL	Verification	Descent Sequence	Level 1 & General SA
ACT-R / IMPRINT	None	Weather System	Level 2
SAMPLE	None	Fighter Air-to-Air	Level 1, 2, 3
TOPAZ	Verification	Runway Incursion	General SA
Linear Regression Model	Validation	Fighter Intercept	General SA
Micro Saint	Validation	Fighter Mission	General SA
IDM / RPD	None	None	Level 1, 2, 3
PCS	Verification	Descent Errors, PF/PNF	Level 1, 2, 3
CSA MIDAS	Validation	Waypoint Navigation	Level 1, 2, 3
A/SA / SEEV / MIDAS	Verification	Taxiway turn errors, Approach, SVS, Wake Vortex, PF/PNF	Level 1, 2
SEEV only	Validation	Approach, SVS, automated cockpit	Level 1

3.3.9 References for SA Deep Dive

Banbury, S., & Tremblay, S. (2004). *A cognitive approach to situation awareness: theory and application*. Ashgate Pub Limited.

Boehm-Davis, D.A., Holt, R.W., Diez, M., & Hansberger, J.T. (2002). Developing and validating cockpit interventions based on cognitive modeling. In W.D. Gray & C.D. Schunn (Eds.) *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science*, p. 27.

Burdick, M.D., & Shively, R.J. (2000). A full-mission evaluation of a computational model Of situational awareness. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA.

Donnelly, D.M., Noyes, J.M., & Johnson, D.M. (1997). Decision making on the flight deck. *IEE Colloquium on Decision Making and Problem Solving*, pp.3/1-3/4, December 16.

Hooey, B. L., Gore, B. F., Wickens, C. D., Scott-Nash, S., Socash, C., Salud, E., & Foyle, D. C. (2011). Modeling Pilot Situation Awareness. *Human Modelling in Assisted Transportation*, 207-213.

Karikawa, D., Takahashi, M., Ishibashi, A., Wakabayashi, T., & Kitamura, M. (2006). Human-machine system simulation for supporting the design and evaluation of reliable aircraft cockpit interface. *SICE-ICASE International Joint Conference*, pp.55-60, October 18-21.

Keller, J., Lebiere, C., & Shay, R. (2004). Cockpit system situational awareness modeling tool. In *Proceedings of the Human Performance, Situation Awareness and Automation Conference (HPSAA II 2004)*, Daytona Beach, FL.

McNally, B.H. (2005). An approach to human behavior modeling in an air force simulation. *Proceedings of the 2005 Winter Simulation Conference*, pp.5 pp., December.

See, J.E., & Vidulich, M.A. (1998). Computer modeling of operator mental workload and situational awareness in simulated air-to-ground combat: An assessment of predictive validity. *The International Journal of Aviation Psychology, 8*(4), 351-375.

Shively, R. J., Brickner, M., & Silbiger, J. (1997). A computational model of situational awareness instantiated in MIDAS. *Proceedings of the Ninth International Symposium on Aviation Psychology*, Columbus, Ohio.

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

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. In the *9th USA/Europe ATM R&D seminar*.

Svensson, E.A.I., & Wilson, G.F. (2002). Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. *The International Journal of Aviation Psychology*, *12*(1), 95-110.

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

Wickens, C. D., Hooey, B. L., Gore, B. F., Sebok, A., & Koenicke, C. S. (2009). Identifying black swans in nextgen: predicting human performance in off-nominal conditions. *Human Factors, 51*, 638-651.

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. Chapter 9 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 213-242.

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.

Zacharias, G. L., Miao, A. X., Illgen, C., Yara, J. M., & Siouris, G. M. (1996). SAMPLE: Situation awareness model for pilot in-the-loop evaluation. *Final Report R*, *95192*.

3.4 Pilot-Automation Interaction Models

3.4.1 Introduction

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 air traffic control (ATC), and monitoring weather and wake vortex conditions, automation is expected to provide pilots with the support needed to perform these tasks.

The information-rich flight decks of the future will require theoretically based design and evaluation methods to ensure that automation is successfully integrated on the flight deck.

Current-day systems show vulnerabilities to the effects of design errors in automation and pilot-automation interaction. This trend is only expected to increase during the transition to NextGen operations. Because the FMS has been the host of much cockpit automation over the past several decades, it will continue to evolve in its role of hosting many new flight path procedures such as those involved in 4D navigation, and in separation responsibilities.

In analyzing the pilot performance models that address pilot-automation interaction, 24 papers were initially reviewed. To be included as a pilot-automation interaction model, there had to be a model of some sort that made specific predictions about pilot performance in automated systems. The model had to provide a way of distinguishing between different types of flight deck automation based on predicted pilot performance. It is possible that other types of models (e.g., workload, error, multitasking) would also be appropriate for evaluating automation, but only those models that were specifically used to address automation are included here. Using these criteria, 16 of the 24 papers were considered relevant for, and included in, the pilot-automation interaction (PAI) deep dive. These papers addressed PAI in a variety of contexts.

3.4.2 Defining a Specific Focus for Pilot-Automation Interaction

Since "human-automation interaction on the flight deck" is a tremendously broad area, the modeling efforts evaluated specific aspects of the domain. Many models focused on a single piece of equipment, several focused on a particular phase of flight, and some were focused exclusively on a particular task.

Many (11 of 16) of the efforts specifically addressed the flight management system (FMS). This is a complex piece of flight deck automation, so it is not surprising that many modeling efforts specifically considered the FMS. Some of the efforts addressed keystroke-level interaction, some evaluated cognitive errors in action sequences, and some predicted failure to notice changes.

Several modeling efforts (5 of 16) specifically evaluated pilot performance at the top of descent or in the arrival / approach phase. These are understood to be the highest workload (and therefore most safety-critical) phases of flight. Many of the models or tools could easily be applied to different phases of flight.

In addition, some models evaluated very specific task sequences. John et al. (2009) and Lüdtke and Osterloh (2010) investigated highly specific procedures. The narrow focus on a limited set of tasks allowed them to study the factors contributing to cognitive errors in great detail. Models by John et al. (2009) evaluated a three-step procedure for entering a landing speed into the Control Display Unit (CDU), and by Lüdtke and Osterloh (2010) evaluated a procedure for changing the flight plan during arrival/approach. Both of these studies predicted cognitive errors in the relatively focused procedural sequences.

3.4.3 Design Tools

Some papers described design tools that were developed to compare different automation designs in terms of their predicted ability to support pilot performance. One effort (Gonzales-Calleros,

Vanderdonckt, Lüdtke, & Osterloh, 2010) evaluated the FMS interface design in terms of its adherence to human factors standards such as font type and color contrast between text and background. The paper outlined an approach for including cognitive models of pilot performance, and described a number of potential benefits, but the model was not actually integrated with the evaluation tool.

Another effort also developed an automation design advisor, called the Automation Design Advisor Tool (ADAT; Sebok et al., 2012) to evaluate and compare potential FMS designs. This effort included multiple analytic models to assess design quality based on human factors principles (e.g., layout of information – providing critical or frequently used information in a readily-accessible region; automation complexity – addressing the number and interaction of modes, and the feedback presented to the pilot regarding current and future status). The analytic models, referred to as modules, evaluated the automation 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, and 6) complexity of procedures necessary to program the FMS, based in part on Sherry et al (2002). ADAT included SEEV and N-SEEV attention models (Wickens & McCarley, 2008; see 3.2.4, 3.3.6 and 3.4.4 below), to predict pilot scanning behavior and the noticeability of FMS information changes. FMS designs are evaluated and compared in terms of their adherence to human factors design principles (through the analytic models) and for their ability to support pilot noticing of mode changes (through the SEEV and N-SEEV models).

The ADAT project included several model verification and validation efforts. When each analytic model was developed, aviation SMEs were consulted to provide feedback on the appropriateness of predictions for a given set of inputs. If the SMEs identified discrepancies between model predictions and real-world experience, model updates were identified and made. Four different reviewers (a professional pilot, an avionics systems designer, a human factors professional, and a graduate student in human factors engineering) performed ADAT-based assessments of flight deck FMS designs. The research team compared the results for a "common sense" evaluation. The most advanced FMS design, Boeing's Flight Deck of the Future (FDOF, Mumaw, Boorman, & Prada, 2006), was rated as best design. ADAT correctly identified known deficiencies in FMS design (e.g., complex mode interactions, lack of feedback about mode operation). In addition, the team compared ADAT scores for assessing procedures in the FDOF and a conventional style FMS. As part of this validation effort, the ADAT-generated procedures scores (from Module 6) were compared with empirical findings of a study that compared procedural learning on a conventional FMS with the FDOF (Mumaw, Boorman, & Prada, 2006). The study revealed that procedural learning was better with the FDOF than with the conventional FMS. In fact, the improvement was approximately 33%. Similarly, the difference in ADAT-predicted procedure scores was 27% higher for the FDOF than for the conventional FMS (Sebok et al., 2012). It was assumed that the difference in procedural learning (observed by Mumaw, Boorman, & Prada, 2006) would correspond to differences in predicted procedural complexity (predicted by ADAT).

Finally, the CogTool design evaluation software (John et al., 2009) allows a designer to create a "use-case storyboard" with a graphical user interface, and predict time to complete task or errors made. In this particular effort, the team focused solely on a three-step procedure of

programming a landing speed into the CDU. They made a series of step-wise adjustments to the underlying ACT-R cognitive model, described below, until the predicted error rates closely approximated those of actual pilot trainees learning to use the FMS.

3.4.4 Predicting Performance Based on Attention / Noticing and Visual Scanning

Boehm-Davis et al. (2002) used an ACT-R model to predict pilot noticing of automation mode changes. ACT-R is a cognitive architecture that attempts to model cognition through goal-directed behavior and a series of "if-then" rules (see page 45 and section 4,2 for more details). The model predicted that pilots were more likely to fail to notice mode changes when the changes were initiated by the automation, rather than the pilot. The authors note that similar trends were observed in previously gathered empirical data.

In the ADAT effort, Sebok et al. (2012) modeled pilot scanning and noticing using the SEEV and N-SEEV models (Wickens & McCarley, 2008). These models predict that visual attention is driven by bottom-up factors of display ***salience*** and ***effort*** (distance from the current viewpoint or need to page among views in a multifunction display) involved in viewing the display, and the top-down factors of ***expectancy*** (more-frequently changing displays will be viewed more often) and ***value*** or importance of the display. Similarly, noticing a change in the visual environment occurs due to these factors. Using input parameters regarding an FMS design (e.g., the location of a change on the flight deck, the salience of the change, if it is always visible or retrieved by a keypress, the workload at the time of the change, the frequency with which the change occurs, and the importance of the change), ADAT (Sebok et al., 2012) predicts the probability that pilots will notice the change and (if noticed) the average predicted time required to notice the change. The SEEV and N-SEEV components of ADAT were not directly validated within the ADAT validation effort. However the SEEV and N-SEEV models, both included in ADAT, have been empirically validated in previous efforts (Wickens, McCarley, Alexander, Thomas, Ambinder & Zheng, 2008; Steelman-Allen, McCarley, & Wickens, 2011). In each of these efforts, model predictions in aviation flight deck contexts, including a high fidelity Boeing 747-400 simulator (Sarter, Mumaw, & Wickens, 2007) were found to predict empirical data of scanning and noticing behavior with correlations above 0.60.

The CASCaS (Cognitive Architecture for Safety Critical Task Simulation) model has been used to predict visual scanning behavior, dwell times in areas of interest (AOIs), and the time required to notice specific visual indications in the cruise and approach phases of flight (Lüdtke, Osterloh, & Frische, 2012). CASCaS, like ACT-R, is a cognitive architecture, which provides a structure and set of rules for simulating human cognition. This effort evaluated pilot performance while interacting with the airborne human machine interface (AHMI; an advanced FMS) in the cruise, approach, and the final approach phases of flight.

The CASCaS model predicted pilot scanning behavior across the phases of flight, and the authors indicated that the overall correlation between model predictions and empirical data was high (r^2=0.85). For specific aspects of model predictions, the average dwell times on the AHMI for cruise, approach, and final approach were 3.2s, 3.2s and 1.7s. Empirical data for those same three phases were 5.7s, 4.9s, and 1.8s. Finally, the average noticing times for visual indications in the cruise and approach changes were compared. The model predicted that average noticing

times would be approximately 1 s in each phase. Empirical data showed that noticing was faster in the cruise phase, with average noticing times being 0.8 s compared with 1.2s in the approach phase. While these results indicate that the model does a reasonable job of approximating pilot behavior, the validations are for highly specific tasks or visual areas, in the context of a much larger set of tasks performed and areas viewed.

In another CASCaS effort (Lüdtke, Osterloh, Mioch, Rister, & Looije, 2009), the CASCaS model predicts cognitive errors such as Learned Carelessness and Cognitive Lockup. Learned Carelessness occurs when pilots routinely perform procedures with multiple steps included to ensure safety criteria are met. If these steps typically do not identify any safety concerns, pilots learn that they improve efficiency by skipping these steps, and they will generally not be working in an unsafe condition. The problem is that sometimes these unsafe conditions do exist, and, by skipping the steps that would allow them to identify the condition, pilots sacrifice safety for efficiency. Another error more directly related to attention is Cognitive Lockup, also known as attention capture. In conditions where an off-nominal event occurs, pilot attention may be drawn to the event and away from routine monitoring tasks (e.g., the Everglades accident, where pilots focused on a faulty landing gear indication and failed to notice that the autopilot had disengaged as the plane descended into terrain).

In two separate efforts, the CASCaS model was implemented, run and compared with empirical pilot performance. The first effort (Lüdtke & Osterloh, 2010) predicted pilot behavior, specifically learned carelessness, in a flight re-planning procedure. The situation of interest was that the re-planning required the pilot to perform a set of verifications of the proposed new route. These required the pilot to verify, through a series of button presses and display observations, that the new route's lateral and vertical trajectories were acceptable. During numerous model runs, the vertical route was acceptable. Over time, the pilot would learn that s/he could skip the vertical route check and accept the new trajectory without problem. However, in two conditions (run 1 and run 24), a problem did exist that could only be identified by checking the vertical view. The model was run 24 times. In the first run, a problem existed in the vertical plan, and the model predicted that the pilot noticed it. Over the next 22 runs, the pilot could ignore the vertical plan, but on the 24th run, another error occurred and had to be noticed in the vertical view. The model predicted that by run 13, the pilot would begin neglecting the checks and would fail to notice the problem in run 24 (i.e., learned carelessness).

The empirical results were similar, in that the subject did notice the first problem, and continued checking the vertical view until trial 12. In contrast to model predictions, the pilot subject showed more erratic behavior. He resumed looking at the vertical route in some later trials. Interestingly, the pilot opened the vertical view page in trial 24, but failed to notice the discrepancy in the flight plan.

Based on reviewing the verbalizations gathered in the experiment, the researchers decided that, in trials 16-17, and 19-21, the participant received an indication that the new plan included a potential timing constraint violation. This cued him to investigate further and see if the route included vertical path violations. The "checking" behavior, once reinitiated, persisted over several trials. This hypothesis, that the pilot was checking for timing constraint concerns rather

than vertical flight path issues, was confirmed by the pilot's failure to notice the vertical path violation (despite having accessed the appropriate display).

The authors updated the model to include "contextual factors" (specifically, strengthening or inhibiting associations between elements in memory). The updated model then correctly predicted (in 23 of the 24 trials) when the pilot checked or neglected to consult the vertical view.

Another effort identifies a high-level attention model using SOAR (State, Operator And Result; Laird, 2008). Like both ACT-R and CASCaS, Soar is a cognitive architecture. The Uijtde Haag, Duan, Schnell et al (2011) effort briefly outlines an approach for combining modeling and simulation tools to predict pilot performance when using hazard and integrity monitoring (HIM) and integrated alerting and notification (IAN) systems on the flight deck. The model can be used to compare different potential system designs. The primary inputs to the modeled pilot are through visual and auditory perception, so attention and noticing are the first step to a more complex cognitive model. Unfortunately the reference to this research project was a presentation, so the details of the modeling effort were sketchy. The authors indicated that a final report had been delivered to the customer (NASA), but the report had not yet been approved for distribution. No validation data are provided.

Polson and Javaux (2001) present a model, discussed in more detail in the deep dive section on workload and multitasking models (3.2.6), and also heavily related to visual attention models, that predicts why pilots do not often scan the flight mode annunciators, a major issue in FMS monitoring. They apply a GOMS 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, and Tautz (1999).

D-OMAR models were developed to predict pilot scanning behavior in different flight deck display configurations (Deutsch & Pew, 2004). One was a baseline, current-day style flight deck, one included a synthetic vision system (SVS) and one included an enhanced SVS (an SVS with primary flight display [PFD] information included, and the PFD removed). The first two conditions had been evaluated empirically, and the D-OMAR predictions were compared with empirical performance. The performance measure was "percent dwell time on the out-the-window (OTW) view, SVS, PFD, and NAV displays." A qualitative comparison of model predictions with empirical data indicated a generally good agreement, but it should be noted that there was substantial inter-subject variability. One key finding was that the presence of the SVS altered the pilots' scanning behavior, so they spent less time viewing the NAV and PFD displays when an SVS was present. With the enhanced SVS, the model predicted that pilot scanning behavior would include "normal" NAV sampling (where the NAV was viewed as frequently as it had been before the SVS was added). No empirical evaluation was performed to validate this particular condition. The researchers indicated that the results were sensible, given that the SVS also included PFD data.

3.4.5 Predicting Performance based on Time to Complete Tasks

CASCaS, described above, was also used to predict time to complete tasks (Lüdtke, Osterloh, & Frische, 2012). The model predicted the time required to handle an uplink from air traffic control (ATC) in the cruise and approach phases of flight. 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 the same conditions revealed that pilots were "faster [to uplink] in the approach phase than in cruise" (but no numbers were provided regarding the magnitude). Discussion with pilot SMEs provided insights into the discrepancy or reversal between model predictions and data. The SMEs indicated that pilots typically have to work faster in the approach phase, 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, and is 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' MIDAS (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 included information on equipment that was monitored and accessed (e.g., displays and controls) when interacting with the three different types of automation, possible interruptions, decision rules, and an algorithm for estimating the time required to make a decision (based on strategy and number of attributes considered). The 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 if it was given via datalink.

To validate the model predictions, Pisanich and Corker conducted an experiment using four 2-pilot crews in a Boeing 747-400 simulator. They collected data on actual pilot performance regarding mode selection. They compared the 'Air' MIDAS model-predicted data and empirical data using three techniques: 1) compared all model predictions with all empirical predictions, 2) compared one model run against all model runs, and 3) compared all empirical data against a single (randomly chosen) model run. By running t-tests of these comparisons, they found no significant difference between the model data and the empirical data. However t-tests are of questionable value in model validation when they are used to assume validity to the extent that model predictions and HITL data show "no significant difference" between them. This is because confirming the null hypothesis in this way does not take any account of the possible low statistical power that may characterize the HITL data. Stated another way, if this power is low,

then it is trivial to show that a HITL data point "does not differ" from a model predictions (using statistics as a criterion), even if the mean value of the predicted point may be quite different.

3.4.6 Predicting Performance based on Workload

Some models predicted pilot performance when using automation based on workload (for more information, see the workload and multitasking deep dive Section 3.2). Gil et al., 2009, (also see Gil & Kaber, 2012) used E-GOMS to model pilot performance when working with a flight control panel (FCP), a 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 need to remember, and the number of information transactions that occur. 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 (the complexity indices varied across automation types). They then conducted a between-subjects human-in-the-loop experiment (for automation types), and 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, as shown in table 3.4.1 below.

Table 3.4.1: Correlations between model predictions and empirical data for complexity indices

Pilot performance measures:	*Heart rate*	*NASA-TLX*	*Vertical flight path deviations*	*Lateral flight path deviations*
Model predictions:				
Number of submethods	$r = 0.928$ $p \leq 0.05$	(positive, but not significant)	$r = 0.978$ $p = 0.008$	No significant correlation
Number of steps	$r = 0.829$ $p \leq 0.05$	(positive, but not significant)	$r = 0.886$ $p = 0.019$	No significant correlation
Chunks of information	$r = 0.928$ $p \leq 0.05$	(positive, but not significant)	$r = 0.978$ $p = 0.008$	No significant correlation
Number of information transactions	$r = 0.883$ $p \leq 0.05$	No significant correlation	$r = 0.971$ $p = 0.001$	No significant correlation

3.4.7 Predicting Performance based on Automation-Induced Errors

Two primary modeling architectures are used to predict pilot error: CASCaS and ACT-R (Adaptive Control of Thought – Rational). These are both cognitive architectures, or frameworks that attempt to model human cognition in terms of noticing information, processing information in working memory, accessing long-term memory, and making decisions.

CASCaS (Lüdtke, Osterloh, & Frische, 2012; Lüdtke & Osterloh, 2010; Lüdtke, Osterloh, Mioch, Rister, & Looije, 2009) has been used to predict pilot error in terms of learned

carelessness and cognitive lockup. Because these factors, while errors, are specifically related to attention and noticing behaviors, the CASCaS papers are summarized in Section3.4.4 , above.

CogTool (John et al., 2009) is based on ACT-R code, and can model the time to complete tasks, errors made on task steps, and failure to complete task steps. In their modeling effort, the authors identified three tasks associated with entering a landing speed into the CDU. They ran the model to predict errors, and performed a series of iterations to improve model predictions. The three tasks were sequential; the first had to be completed before the next could be performed. The CogTool model 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 being accessed was developed to represent a college student's knowledge, not the specialized knowledge that a pilot would possess. They then accessed an aviation-specific corpus of terms, and the success rate jumped to 10 percent for the first task only. A series of other changes were implemented, to account for the specialized knowledge that pilots possess, 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.

In addition, as discussed also in Section 3.2.6, Schoppek and Boehm-Davis (2004) used ACT-R to create a model (ACT-Fly) to model pilot awareness, cognition, and errors. They, like the Pisanich and Corker (1995) study, evaluated pilot use of automation at the end of the cruise phase of flight until the initial approach fix. They developed a model to predict when pilots would choose a more automated mode (VNAV, in which a preprogrammed flight plan provides reference values for the flight) or a less-automated mode (FLCH and V/S, which require the pilot to provide reference values). The authors implemented a relatively simple heuristic for selecting the mode: if the ratio of the horizontal distance to vertical distance to the waypoint is greater than 250, they will chose a less-automated mode. The researchers ran the model 6 times for each scenario. Both scenarios were a descent through 6 waypoints. In one scenario, there were no ATC clearances, but in the other, a clearance was issued shortly after the top of descent for waypoint 3. The model predicted that pilots would chose VNAV for all but the last leg of the flight for the "no ATC clearance" condition, and that they would chose FLCH for the "clearance" condition. Further, in 12 runs, the model predicted 4 errors of omission, 1 of which the modeled pilot never recovered. In the other 3 error conditions, the pilots failed to resume a deferred action.

Empirical data were gathered from 5 commercial pilots using a desktop simulator with keyboard and mouse controls. The 5 pilots participated in 2 scenarios each. Scenario 1 should have been handled using the V/S or FLCH mode, and scenario 2 should have been handled with the VNAV mode. Only 1 subject in scenario 1 used the V/S mode. Two subjects incorrectly used the VNAV mode exclusively for scenario 1, and the remaining 2 subjects switched frequently among different modes (e.g., FLCH, VNAV, and V/S). This suggests a 20% agreement between model predictions and empirical data for scenario 1. Scenario 2 revealed better (60%) agreement, with 3 of the subjects using the model-predicted VNAV mode. The other 2 subjects switched frequently among modes.

The actual pilots performed surprisingly poorly, and made many more mode changes than the model predicted. Further, they missed crossing restrictions and made poor decisions about initiating an early descent. The researchers attributed the pilots' poor performance to the fact that they work in 2-person crews, yet the experiment was designed for a single pilot. In addition, the controls (keyboard and mouse rather than a control yoke and rudder pedals) were a potential interference.

The authors conclude that ACT-Fly had certain strengths and limitations. It could model errors of omission and commission, and it could be used to evaluate pilot performance when working with different procedures. The model did not adequately address mode surprise, and modeled pilots, once they made an error, were much poorer than humans are at recovering from it. Finally, the model did not address pattern recognition, which is commonly used by pilots in interpreting situations and planning responses. Rather, ACT-R relies only on rules, and therefore fails to capture the complexity of human cognition.

Gil et al (2009) indicate that their E-GOMS based model, discussed in Section 3.4.6, can be used to predict error. When the number of chunks to be held in working memory exceeds 5, the model predicts an increased likelihood of pilot error, based on limitations of working memory (Miller, 1956). No error predictions were made in the paper, and the empirical validation did not include error data.

3.4.8 Trust in Automation

Another aspect of pilot-automation interaction is trust in automation. Raeth and Reising (1997) developed a model based on Lee and Moray's (1992) regression equation that attempts to predict the degree of trust an operator will have for the automation system being considered. The Lee and Moray (1992) model, developed using data gathered from operators working with a process control simulation, predicts that "current level of trust" is based on the previous level of trust the operator had, the current system performance, previous system performance, the occurrence of a fault, and previous occurrences of faults. The authors conducted an empirical study used to provide the basis for the regression equation. Raeth and Reising (1997) updated this model and gathered model-prediction data for an aviation setting. Their updated trust model included a history of automation performance over time, continuous (rather than discrete) fault ranges, and added a parameter to account for "danger." They also developed two different models to predict the trust level of an expert versus a novice pilot. The difference between these equations is that a novice pilot has a more volatile sense of trust; s/he is more likely to base trust on the immediate previous experiences. In contrast, an experienced pilot has a deeper understanding of the automation, and his/her trust is less swayed by faults or recent events. Raeth and Reising ran their model for a particular flight combat scenario (including dangerous events) and found that the results of trust changes through the scenario for the expert and novice pilots generally matched the researchers' qualitative expectations.

3.4.9 Adaptive Automation

Another important issue in automated systems, with potential applications in NextGen airspace, is the use of adaptive automation. These systems attempt to infer when an operator or pilot is in need of additional support, typically based on high workload or poor task performance, and to

provide this assistance through intervening automation. None of the sources reviewed in this effort identified models that specifically address pilot performance while using adaptive automation, but several papers ***did address ways to predict when the pilot would be in need of assistance***. These papers typically included visual attention or workload models (see 3.2).

Donath and Schulte (2009) describe a study (with one pilot as a subject) in which eye movements were measured during high and low workload tasks in a flight simulator. The intent of the effort was to develop adaptive automation, and this preliminary investigation was to see if eye movements could provide an indication of high workload, and thus serve as a prompting mechanism for automation to "step in and offer assistance." The results indicated that, in high workload situations, eye movements did indeed change; the participant viewed fewer visual areas of interest, and exhibited tunnel vision. No model was developed or proposed, and the authors stated that individual "calibrations" would need to be performed for each pilot.

Bzostek, Small, Bagnall, and Walters (2005) developed a model to predict pilot visual and auditory workload during flight, based on the types of indications presented on the flight deck. The long-term goal of this effort was to develop adaptive information presentation systems that would give pilots critical information in the modality (visual or auditory) that they would be most likely to notice. The authors' modeling indicated potential benefits of context-sensitive information presentation. "Correct modality" indications were noticed more than 20 percent faster than baseline conditions. There was no empirical validation.

Similarly, Schurr (2011) developed an adaptive automation system for presenting different levels of alert warnings, for both external hazards (e.g., traffic conflicts) and internal hazards (e.g, engine conditions). Fundamental to the system was a model of what the pilot needed to know (e.g., be alerted about by automation), and in what priority, given multiple warnings. Such a model included an inference of pilot state and workload, as is essential to all effective adaptive automation systems. This system, and hence the model contained within, however, was not validated.

The ALARMS (the ALerting and Reasoning Management System; Carlin et al., 2010) project included developing a pilot model for predicting when additional stages of automation are needed for an integrated alerting system. ALARMS used a workload-based pilot performance model to support prioritization of alarms and instantiate different levels of automation. The model included three types of pilot workload: mental effort, task demands, and ongoing task performance. These are input by the ALARMS user, as one of three possibilities: low, medium, or high. The ALARMS user also inputs other data to characterize a scenario (hazard sources, and type of indication). ALARMS predicts when different stages of automation will need to be invoked, and provides displays to support pilots in those scenarios. No empirical validation was performed. ALARMS can be used for the en route, approach, and landing phases of flight. Substantial efforts were made to identify relevant scenarios, and to develop information presentation formats. The researchers performed a work domain analysis, and cognitive task analyses, working with aviation SMEs and using aviation procedures. These efforts went into developing the models and tools; they were not actually used after the fact to verify that the predictions were reasonable.

3.4.10 Summary

There are many ways to model pilot-automation interaction and predict performance on the flight deck. These can include 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 due to learning certain patterns in the automation. Because there are so many factors that can have an influence, it is difficult to capture all in a single model. The ADAT software does integrate several process models applicable to the FMS in one set of software. However, to date, the CASCaS effort (Lüdtke, Osterloh, & Frische, 2012) appears to be the most comprehensive type of pilot performance model, addressing attention, interaction, and errors.

It is important to note that certain critical aspects of pilot-automation interaction beyond the FMS have yet to be well addressed by computational modeling. Three of the most prominent are decision aiding, alerting responses, and adaptive automation. In decision aiding category, we might include models that predict the pilot's compliance with intelligent decision aids regarding route diversion around weather or engine failure diagnostics. In the alerting responses category, we might include application of signal detection models to predict pilot's degree of compliance with trajectory collision alerts and resolution advisories (e.g., TCAS). As described above, adaptive automation models are lacking.

3.4.11 Statistics of Validation of PAI Models

Validation Data

Of the 16 papers reviewed for this deep dive, 8 included a validation effort, although only three reported correlation data. These validation efforts have been described above, when the model was introduced. Typically the validations were limited in scope, to address only specific aspects of model predictions.

- Deutsch & Pew (2004) – qualitative comparisons of scanning behavior among areas of interest on a flight deck
- Gil, Kaber, et al (2009) – correlations between workload measures and complexity indices
- Lüdtke & Osterloh (2010) – comparison of performance in 24 trials (when did pilots fail to verify or fail to notice?)
- Lüdtke, Osterloh & Frische (2012) – validation of a very limited set of parameters
- Manton & Hughes (1990) – regression equation to predict interaction time
- Pisanich & Corker (1995) – "no significant difference" between model predictions and empirical performance
- Schoppek & Boehm-Davis (2004) – qualitative comparison of mode changes
- Sebok, Wickens, et al (2012) – procedures module, SEEV and N-SEEV models. Comparisons between ADAT predictions and performance in comparing new and old designs.

Validations were often qualitative, comparing the pattern of results with the general pattern of model predictions (e.g., Lüdtke & Osterloh, 2010 – *when did pilots fail to verify?*; Schoppek & Boehm-Davis, 2004 – *when did pilots switch mode?*). Further, one validation effort (Pisanich &

Corker, 1995) used a t-test to show that empirical results were "not significantly different from" model predictions.

Verification Efforts

Many modeling efforts included a verification component. To complete the verification process, authors performed task analyses, observed expert performance, gathered empirical data to inform model development. In addition, most authors also consulted with aviation SMEs to perform a "reality check" of results, to identify and explain differences between model predictions and actual (empirically-gathered) or expected (based on operational experience) performance. Further, researchers typically attempted to perform their own "common-sense" evaluation of results.

3.4.12 References Included in the Pilot-Automation Interaction Deep Dive

Boehm-Davis, D.A., Holt, R.W., Diez, M., & Hansberger, J.T. (2002). Developing and validating cockpit interventions based on cognitive modeling. In W.D. Gray & C.D. Schunn (Eds.) *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science*, p. 27.

Bzostek, J., Small, R., Bagnall, T., & Walters, B. (2006, October). Intelligent multimodal signal adaptation system (IMSAS). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 11, pp. 1170-1174). SAGE Publications.

Carlin, A.S., Alexander, A.L., & Schurr, N. (2010). Modeling pilot state in next generation aircraft alert systems. Aptima, Inc.

Deutsch, S., & Pew, R. (2004). Examining new flight deck technology using human performance modeling. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA.

Gil, G., Kaber, D., Kim, S., Kaufmann, K., Veil, T. & Picciano, P. (2009). Modeling pilot cognitive behavior for predicting performance and workload effects of cockpit automation. *Proceedings of the 2009 International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, 124-129.

Gonzales-Calleros, J., Vanderdonckt, J., Lüdtke, A. & Osterloh, J.P. (2010). Towards model-based AHMI development. *EICS '10*. June 21-23, Berlin, Germany.

John, B.E., Blackmon, M.H., Polson, P.G., Fennell, K., & Teo, L. (2009). Rapid theory prototyping: An example of an aviation task. In the *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. *Paper presented at the 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. (2010). Cognitive modelling of pilot errors and error recovery in flight management tasks. In the *Proceedings of* 7th IFIP

WG 13.5 Working Conference, HESSD 2009, Brussels, Belgium, September 23-25, 2009, Revised Selected Papers, (pp 54-67).

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

Pisanich, G. M., & Corker, K. M. (1995, April). A predictive model of flight crew performance in automated air traffic control and flight management operations. In *Proceedings of the 8th international symposium on aviation psychology* (pp. 335-340).

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

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.

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., Quesada, S., Socash, C., Anthony, B. (2012). The Automation Design Advisor Tool (ADAT): Development and Validation of a Model-Based Tool to Support Flight Deck Automation Design for NextGen Operations. *Human Factors and Ergonomics in Manufacturing and Service Industries*, *22*(5), 378-394.

Uijtde Haag, M., Duan, P., Schnell, T., Cover, M., Anderson, N., Snow, M., Etherington, T., Rademaker, R., & Theunissen, E. (2011). *Hazard and Integrity Monitoring and Integrated Alerting and Notification Methods*. Presentation delivered to the 2011 NASA Aviation Safety Technical Meeting in St. Louis, MO.

3.5. Roles and Responsibilities (R&R) Models

3.5.1 Introduction

Roles and responsibilities (R&R) have been modeled from two perspectives. One is the shift in R&R from ground (ATC) to the air (pilot) brought about by many NextGen considerations and ConOps (or their predecessors in Free Flight). The other is the shift or distribution of responsibilities between pilot flying and pilot not flying (or pilot flying PF and pilot monitoring, PM; the terms we used here). We also note three different classes of questions that the models address. The first are those in which the model simply predicts performance of "the crew" as a whole, but does not attempt (or the validation effort does not portray) the separate behavior and cognition of PF and PM. The second are those in which separate representations of PF and PM are predicted by the model. The third, and probably most relevant to the R&R concept (but least available) are models where a change in R&R is modeled; for example, as above, a shift from ground to air for traffic separation management, or an added responsibility given to PM or PF for some duty, imposed by NextGen. The following describes four major clusters of modeling efforts.

3.5.2 MIDAS Efforts

Two studies were done with an earlier version of the MIDAS software ('Air' MIDAS, MIDAS v1). Pisanich and Corker (1995) describe two 'Air' MIDAS applications to crew performance (without distinctly describing R&R for PF and PM). The first of these, described also in 3.4.7, examined the latency of the crew's decision to select different automation modes prior to the top of descent, as a function of the level of automation of the mode, and the time remaining to TOD. The authors report a qualitative validation that the model "behaved" in a way that was consistent with observations of pilots carrying out a corresponding LOFT scenario. In the second effort described in the same paper, the authors programmed MIDAS to process and accept the advice of the CTAS Descent Advisor, an early form of automated decision aiding, such as that being considered in current NextGen plans. Predicted times to accept were modeled, and parameters adjusted based upon four crews flying a corresponding simulated approach. Then the 'Air' MIDAS model predicted times were compared against four additional crews, in a "split-halves" validation. Predicted and obtained times were found not to differ from each other, based upon a t-test of equivalence. However, we have noted earlier that this is a questionable approach for model validation and the authors do not provide information on the actual acceptance times (for model and for pilots), so that the degree of correspondence cannot be established.

In Corker and Pisanich (1998) the authors use 'Air' MIDAS to generate a prediction of the crew's use of the CDTI (cockpit display of traffic information) to initiate a break-off maneuver when separation from traffic aircraft went below a minimum. Different levels of traffic density were also varied. The authors again compare 'Air' MIDAS predictions with obtained data, reporting no differences (via a t-test), but provide no further data regarding actual predicted and obtained times. Thus, as with Pisanich and Corker (1995) this is a weak form of validation. Both papers however do a good job of explaining the workings of 'Air' MIDAS.

Three MIDAS studies of R&R (using MIDAS v5), have provided a better focus on the actual differences in R&R between PF and PM. Gore, Hooey, Haan, Bakowski and Mahlsted (2011) used MIDAS v5 to model the descent with both a conventional and an "augmented" display suite, including a highway in the sky (HITS) display for the pilot on a HUD, and an advanced navigational planning display. This effort, described also in the SA deep dive (3.4.6), examined the differences in PF and PM SA, for aviation awareness, navigational awareness and hazard awareness that were created by the implementation of the augmented display suite and differentiated pilot roles and responsibilities. In one condition, both pilots adhered to the conventional task hierarchy (aviate, [separate], navigate, communicate, and systems management. In the second, augmented, condition, the PF emphasized the tasks of Aviate and Separate (from immediate hazards) while the PM emphasized the tasks of Navigate and Separate (from global hazards). These differences were mediated in the Situation Awareness model by changes in task responsibilities and by predicted changes in scan pattern. No validation data were provided.

In a second study with the MIDAS v5, Gore, Hooey, Mahlstedt and Foyle (2013) examine how the shift in R&R for traffic separation management, from ground to air would change predictions. The workload predictions and traffic maneuver time predictions were discussed in the context of the workload deep dive. Here we highlight that, using the SEEV model (discussed in other deep dives), the authors present predictive data on this responsibility-induced change in

scanning (i.e., allocation of visual attention) across three displays (Navigation, primary flight, out-the-window), as differentiated between PF and PM.

In a third study (Gore, Hooey, Mahlstedt & Foyle, 2013), two different R&R configurations were defined on an RNP approach. In one, both PM and PF were equally responsible for all aspects of the flight (aviating, navigating, and hazard separation). In the other, responsibilities were distributed so PF was primarily responsible for aviating, and PM was primarily responsible for the other two meta-tasks. MIDAS was run under both conditions, and data for channel workload (based upon the VACP model, see workload deep dive), for scanning (based on the SEEV model), and for detection of an off-nominal event (an RNAV violation alert) were predicted and found to produce plausible effects (e.g., verification). No validation was provided.

3.5.3 Distributed Workload

Lyall and Cooper (1992), as summarized in the workload deep dive (3.2), also focused on using the VACP model to predict the distribution of workload between PF and PM (as with Gore, Hooey, Malhstedt & Foyle, 2013, above), and how this workload distribution changed between the two crew members over two different departure procedures. A similar approach was taken by Stone et al (1987) in a paper also reviewed in the workload deep dive. The advantage of such model predictions is that they can be used to identify workload imbalances, and hence justify shifts in task responsibilities to the less-loaded crew member. Validation data were not apparently collected for these modeling efforts.

3.5.4 Knowledge Based Models

ACT-R was employed by Boehm-Davis Holt, Chong, and Hasberger (2004) to model PF and PM behavior during a top-of-descent period. The model was used to predict the effects of different flight workload levels (imposing an "ATC task", whose nature was undefined in the article) and different levels of skill/knowledge, a feature that ACT-R is ideally suited to predict. ACT-R was augmented in this two-crew scenario to include a communications module; between separate ACT-R models of PF and PM. Primary model outputs were the number of procedural steps completed or omitted, and important distinctions between these for PF and PM were observed across the four conditions of workload and knowledge. There was no validation of these predictions.

Tidhar et al. (1998), present a model called SWARMM (smart whole air mission model), based on a SOAR architecture, which is designed directly to characterize distributed responsibilities and tasks in two-seat fighter aircraft. As such it is only indirectly relevant to the commercial cockpit, because the nature of duty assignments (between flying the aircraft and weapons management) is quite different. There is little presentation of model predictions, and no validations. However the paper may be an important reference for those undertaking R&R models, because of the nature of assumptions made regarding team performance.

3.5.5 Distributed Risk Model

A paper from the NLR (Netherlands Aerospace Research Lab), by Blom, Stroeve and their colleagues was centered around one generic model – TOPAZ (Stroeve, Blom, & Baaker, 2009), also described in the SA deep dive (3.3.3). The model focused on predicting runway incursions resulting from one aircraft failing to stop at an intersection where another was on a take-off run.

(Other applications of TOPAZ by these same authors are treated in the error deep dive section.) The focus of the model was to predict objective risks. Thus this model incorporated three component models of two pilots and a ground controller in their interaction, and hence highlights the different roles and responsibilities of the three agents. Because it was a Monte Carlo simulation model, repeated runs predicted the relative frequency of these very rare events (runway collisions). 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, and the contributions of the different agents to this joint risk likelihood.

While the model only reports verification rather than validation, to some extent this limitation is understandable, given that the extremely low number of such incursions means that capturing reliable numerical estimates of their frequency of occurrence (and hence, reliable targets for model prediction) is very difficult.

3.5.6 Conclusion: R&R Models

A positive feature of these models is that a majority of model verifications have examined the most important and NextGen-relevant kind of verification: of the **change** in R&R brought about by flight deck and/or procedural changes. Most have looked at changes or differences within the flight deck, while only a minority has examined changes between ground and air, highly relevant for NextGen. Finally, we note that most of them have produced predictions that can generalize to NextGen specific changes or procedures. Hence many of these models are quite promising.

3.5.7 Statistics of Validation of R&R Models

Altogether 9 separate modeling efforts, including a total of 12 applications, were identified that focused on roles and responsibilities. Of these, three could be said to be of category 1 (overall crew performance), two of category 2 (different predictions for the players on the team, but no examination of the change in R&R brought about by changes in procedure or flight deck), and seven were of category 3 (predicting such R&R changes).

In terms of validation, only three of the efforts included validation data, and of these, two used inappropriate t-test statistics to assess validation (and did not contain performance data to allow readers to compute a correlation). The third used a qualitative form of validation.

3.5.8 References for the Roles and Responsibilities Model Review

Boehm-Davis, D. A., Holt, R. W., Chong, R., & Hasberger, T. (2004). Using cognitive modeling to understand crew behavior. *Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA.

Corker, K. M. & G. Pisanich (1998). Cognitive performance for multiple operators in complex dynamic airspace systems: Computational representation and empirical analyses. *Proceedings of the Human Factors and Ergonomics Society.* 1: 341-345.

Gore, B.F., Hooey, B.L., Mahlstedt, E., & Foyle, D.C. (2013). *Evaluating nextgen closely spaced parallel approach concepts with validated human performance models flight deck guidelines* (Part 2 of 2). In Human Centered Systems Lab, HCSL Technical Report (HCSL-12-02). Moffett Field, CA: NASA Ames Research Center.

Hooey, B. L., Gore, B. F., Wickens, C. D., Salud, E., Scott-Nash, S., Socash, C., & Foyle, D. C. (2010). Modeling pilot situation awareness. Paper presented at the *Human Modelling of Assisted Technologies Workshop*, Belgirate, Italy.

Lyall, E.A., & Cooper, B., (1992). The impact of trends in complexity in the cockpit on flying skills and aircraft operation. *Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA.

Pisanich, G.M. & Corker, K.M. (1995). A Predictive Model of Flight Crew Performance in Automated Air Traffic Control and Flight Management Operations. Proceedings of the Ohio State 8th *International Symposium on Aviation Psychology*, Columbus, Ohio.

Stone, G., Culick, R., & Gabriel, R. (1987) Use of task timeline analysis to assess crew workload. In A. Roscoe (Ed) *The practical assessment of pilot workload*. NATO AGARDograph #282.

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

Tidhar, G., C. Heinze, & Selvestrel, M. (1998). Flying together: modeling air mission teams. Applied Intelligence: *The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 8*(3): 195-218.

3.6 State of Verification and Validation Efforts Across Deep Dive Models

Table 3.6.1 below summarizes the verification and validation efforts for the individual models reviewed in the deep dive analysis. Each modeling effort is rated along an 8-point scale, with an X provided in the column corresponding to the level of verification or validation performed. The first four points indicate that a verification was performed. The second four points indicate that a validation was performed. The individual numbers provide a ranking of the quality of the verification or validation of the model, and correspond to the following criteria:

1. Used SME-based data to build the model. This includes a cognitive task analysis, work domain analysis, or similar.
2. The researchers performed a reality check of the model predictions – did the results make sense?
3. The researchers had aviation SMEs perform a reality check of model predictions.
4. The researchers worked with SMEs to conduct systematic walkthroughs of results to identify concerns.
5. Empirical data were compared with model predictions for a qualitative validation assessment (did the general trends match?)
6. Correlation data were used, but the subjects were students (or non-pilots) working on a low-fidelity (e.g., desk-top) or non-aviation simulation.

7. Correlation data were used, but either the subjects were not pilots or the simulation was not a high-fidelity aviation simulation.
8. Correlation data were used, subjects were pilots working on a flight simulator or in actual operations.

Note that this table does not distinguish between partial and complete validation efforts. Further, many validation efforts included different degrees of validation. For example, a model that predicts visual scanning, errors, and workload might include qualitative analyses of the error and workload data, but also include a quantitative evaluation of scanning data. For simplicity, the table below indicates the ***highest degree of verification or validation*** performed in the particular study. This can give a somewhat optimistic picture of the state of verification and validation efforts. For any modeling effort that included some level of validation, we assume that a verification will have taken place, but we have not checked a particular box within the four verification columns at the left.

Finally, it should be noted that some modeling efforts did not describe any attempts to obtain SME input. These are indicated with a "- -" in the first column of the table. This lack of SME input could be due to the authors not explaining where they gathered their data, or (in the case of Gore, 2008) because the paper describes an approach to modeling, rather than a specific modeling project with a certain degree of verification or validation. A corresponding table for modeling architectures is presented in Section 4.

Modeling Effort	*Model or Architecture*	*Deep dive in which it was evaluated*	*Verification*				*Validation*			
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Blom, Corker, Stroeve & van der Park (2003)	Air MIDAS	E	X							
Boehm-Davis, Holt, Chong & Hansberger (2004)	ACT-R	R&R				X				
Boehm-Davis, Holt, Diez & Hansberger (2002)	ACT-R	A, SA					X			
Burdick & Schively (2000) / Shively, Brickner & Silbiger (1997)	MIDAS	SA								X
Byrne, Kirlik & Fleetwood (2008)	ACT-R	E					X			
Carlin, Alexander & Schurr (2011)	ALARMS	A	X							
Corker & Pisanich (1998)	Air MIDAS	R&R					X			
Corker, Muraoka, Verma, Jadhav & Gore (2008)	Air MIDAS	E					X			
Deutsch & Pew (2004)	D-OMAR	A					X			
Deutsch & Pew (2008)	D-OMAR	E, WL, MT					X			
Donnelly, Noyes & Johnson (1997)	IDM	SA	X							
Fotta, Nicholson & Byrne (2007)	ACT-R	E	X							
Gil, Kaber, Kim, Kaufmann, Veil & Picciano (2009)	E-GOMSL	A, WL								X
Gonzales-Calleros, Vanderdonckt, Lüdtke & Osterloh (2010)	Usability Advisor	A	--							
Gore & Corker (2000a)	Air MIDAS	WL			X					
Gore & Corker (2000b)	Air MIDAS	WL			X					
Gore (2008)	MIDAS	WL	--							
Gore, Hooey, Mahlstedt & Foyle (2013)	MIDAS	R&R				X				

Modeling Effort	*Model or Architecture*	*Deep dive in which it was evaluated*	*Verification*				*Validation*			
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Gore, Hooey, Socash, Haan, Mahlsted, Bakowski, Gacy, Wickens, Gosakan & Foyle (2011)	MIDAS	WL								X
Hooey, Gore, Wickens, Salud, Scott-Nash, Socash & Foyle (2010)	MIDAS	SA, R&R				X				
John, Blackmon, Polson, Fennell & Teo (2009)	ACT-R	A				X				
Karikawa, Takahashi, Ishibashi, Wakabayashi & Kitamura (2006)	PCS	E, SA					X			
Keller, Lebiere & Shay (2004)	ACT-R	SA			X					
Laudeman & Palmer (1995)	TLAP	WL, MT								X
Lebiere, Archer, Best & Schunk (2008)	ACT-R	E, WL					X			
Lüdtke & Osterloh (2010)	CASCaS	A					X			
Lüdtke, Osterloh & Frische (2012)	CASCaS	A							X	
Lüdtke, Osterloh, Mioch, Rister & Looije (2009)	CASCaS	E, A					X			
Lyall & Cooper (1992)	MRT	R&R, WL, MT		X						
Manton & Hughes (1990)	MRT	A, WL				X				
McNally (2005) / Zacharias, Miao et al (1996)	SOAR	SA		X						
Miller (1998)	MRT	WL		X						
Miller (2001)	ASHRAM	E	--							
Muraoka & Tsuda (2006)	OPSAMS	WL	X							
Nikolic & Sarter (2003)	(ASRS data)	E								X
Parks & Bouceck (1989)	TLAP	WL								X
Pisanich & Corker (1995)	Air MIDAS	A, R&R					X			
Polson & Javaux (2001)	GOMS	A, MT		X						

Modeling Effort	*Model or Architecture*	*Deep dive in which it was evaluated*	*Verification*				*Validation*			
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Raeth & Reising (1997)	(trust / workload allocation)	A		X						
Rickard & Levison (1981)	OCM	WL								X
Riley, Lyall, Cooper & Wiener (1991)	MRT	MT, WL								X
Salmon, Stanton, Young, Harris, Demagalski, Marshall, Waldman & Dekker (2002)	SHERPA	E					X			
Salmon, Stanton, Young, Harris, Demagalski, Marshall, Waldman & Dekker (2003)	SHERPA	E					X			
Sarno & Wickens (1995)	MRT	WL						X		
Schoelles & Gray (2011)	ACT-R	MT		X						
Schoppek & Boehm-Davis (2004)	ACT-R	A, MT					X			
Schurr (2011)	ALARMS	WL	X							
Sebok, Wickens, Sarter, Quesada, Socash & Anthony (2012)	SEEV, N-SEEV, ADAT	A, WL								X
See & Vidulich (1998)	Microsaint	SA, WL							X	
Stanton, Salmon, Harris, Demagalski, Marshall, Waldman & Dekker (2003)	SHERPA	E					X			
Steelman-Allen, McCarley & Wickens (2011)	N-SEEV	WL								X
Stone, Culick & Gabriel (1987)	TLAP	R&R, WL				X				
Stroeve & Blom (2005)	TOPAZ	E		X						
Stroeve, Blom & Bakker (2009)	TOPAZ	E, SA, R&R			X					
Stroeve, Blom & Bakker (2011)	TOPAZ	E, SA			X					

Modeling Effort	*Model or Architecture*	*Deep dive in which it was evaluated*	*Verification*				*Validation*			
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Svensson & Wilson (2002)	(regression)	SA					X			
Tidhar, Heinze & Selvestrel (1998)	SWARMM	R&R	X							
Uijtde Haag, Duan, Schnell, Cover, Anderson, Snow, Etherton, Rademaker & Theunissen (2011)	SOAR	A	--							
Walden & Rouse (1978)	(queueing)	MT					X			
Wickens, Bagnall, Gosakan & Walters (2011)	N-SEEV	MT		X						
Wickens, Goh, Helleberg, Horrey & Talleur (2003)	SEEV	MT, SA							X	
Wickens, Harwood, Segal, Tkalcevic & Sherman (1988)	TLAP, MRT	MT, WL							X	
Wickens, Hooey, Gore, Sebok & Koenecke (2009)	N-SEEV	SA							X	
Wickens, Larish & Contoror (1989)	TLAP, MRT	WL							X	
Wickens, McCarley, Alexander, Thomas, Ambinder & Zheng (2008)	A/SA	E, MT, SA					X			
Wickens, Sebok, Kamienski & Bagnall (2007)	A/SA, SEEV	SA				X				

Notes: A-Automation interaction, E-Error, MT – Multi-task, SA – Situation Awareness, WL – Workload, R&R – Roles and Responsibility

4. Model Architectures

4.1 Overview

Model architectures are defined differently than individual modeling efforts. A model architecture is a framework for creating any number of more specific models to predict hu performance. The architecture can be a framework for representing cognition or attention ACT-R or SEEV) that eventually needs to be implemented within a programming languag an architecture can be a specific modeling tool that allows users to create models (e.g., IMPRINT).

Across our reviews of modeling efforts, through all five deep dives, we have identified sev architectures that were used in multiple modeling efforts, and were often used to model m aspects of pilot performance (e.g., situation awareness, workload). We describe these architectures below. While these descriptions may be partially redundant with the specific effort applications discussed in Section 3, our emphasis here is somewhat different: how d model work, how might it contribute to several different modeling aspects, and how usabl available is it?

Table 4.1 provides a list of key architectures and the aspects of pilot performance that they modeled. These data were compiled from the spreadsheet developed as part of this researc Each modeling effort that could be associated with a specific architecture (e.g., ACT-R, M was identified. The specific aspects of pilot performance that were modeled in these effor also identified and indicated with *'s in the table. Finally, the extent to which the architec had been validated was calculated by taking the total number of modeling efforts (again, i spreadsheet) for a given architecture and dividing that into the total number of validated ef for that architecture.

Table 4.1: Pilot modeling aspects and validation efforts addressed by different major architectures.

Pilot Aspect:	Human-Automation Interaction	Communication	Decision making	Error	Manual Control	Multi-Tasking	Procedures	Roles & Responsibilities	Situation Awareness	Spatial Disorientation	Visual scanning	Workload	Percent of efforts that included at least a partial *Validation*
Architecture													
ACT-R	*	*	*	*	--	*	*	*	*	--	*	*	6/15; 40%
Air MIDAS	*	*	*	*	--	--	*	--	--	--	*	*	6/11; 55%
CASCaS	*	--	--	*	--	*	*	--	--	--	*	--	4/10; 40%
MIDAS	--	*	*	*	--	*	*	*	*	--	*	*	6/11; 55%
MRT	*	--	--	--	--	*	--	*	--	--	--	*	3/8; 38%
OCM	--	--	--	--	*	--	--	--	--	--	*	*	7/9; 78%
SEEV, N-SEEV, A/SA	*	--	*	--	--	*	--	--	*	--	*	--	5/7; 71%
TLAP	--	--	--	--	--	*	--	--	--	--	--	*	4/5; 80%
TOPAZ	--	*	*	*	--	--	*	--	*	--	--	--	1/6; 17%

- MRT – includes W/Index

4.2 Adaptive Control of Thought – Rational (ACT-R)

4.2.1 What Is It?

The Adaptive Control of Thought – Rational (ACT-R) is a unified theory of cognition that integrates theories of attention, cognition, and motor actions. It was developed by Carnegie Mellon University for the Office of Naval Research in 1993, and has been in use and updated multiple times since then. ACT-R is a theory of cognition that has been represented as executable software that is programmed in LISP.

ACT-R provides a framework for representing human cognition. It models cognition through "production rules" or goal-directed behavior that is implemented through a series of "if-then" rules. It includes perceptual inputs and motor outputs.

ACT-R's main components are modules, buffers, and the pattern matcher. ACT-R uses perceptual-motor modules (visual and manual modules) to simulate interaction with the physical environment. Memory modules simulate declarative and procedural memory and allow the modeler to represent an operator accessing different types of information from long-term memory. Declarative memory consists of simple facts. Procedural memory consists of representations for how to perform different tasks.

Buffers are interfaces for each of the ACT-R modules, except the procedural memory module. The contents of each buffer at a given time represent the state of ACT-R at that moment. The pattern matcher uses the buffer contents to identify a relevant schema to select goals and behavioral rules. Cognition is modeled as a succession of these changes as instructed by the pattern matcher.

ACT-R does not have a goal hierarchy, as a pre-established hierarchy would suggest perfect memory on the part of the modeled operator. With the pattern matcher, whatever situation provides the closest match is what changes the state of the system.

Another important aspect of ACT-R is that it models the actual time required for cognitive steps (e.g., retrieving and 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 times, provided as default parameters in ACT-R, are based on psychological theories or empirical research (e.g., 50 msec to retrieve information; Fitts' law for scanning a display or selecting a control). In addition to timing aspects, ACT-R models errors of omission, where a retrieval from memory fails, and errors of commission, where an error occurs due to imperfect matching.

ACT-R 6 version 1.4 [r1261] (the latest version as of August 9, 2012) is available on the website: http://act-r.psy.cmu.edu/actr6/

Inputs: Data from psychology experiments, general assumptions about human cognition, assumptions about a particular domain

Outputs: overt behavior, time to perform the task, accuracy in the task, and whether or not the task was performed

4.2.2 What Has It Been Applied To?

ACT-R has been used in aviation, and in many other domains. These include human-computer interaction, department of defense research, education (in cognitive tutoring systems), and neuropsychology. It has been used to predict the time to complete task sequences (Fleetwood, Lebiere, Archer, Mui, & Gosakan, 2006), if operator error occurs (John et al., 2009), and the type of error (Boehm-Davis et al., 2002).

ACT-R models have been used in more than 700 different scientific publications.
A long list of references is available: http://act-r.psy.cmu.edu/publications/index.php.

4.2.3 Where Can You Get It?

ACT-R is an open-source architecture, available online at:
http://act-r.psy.cmu.edu/

4.2.4 How Usable Is It?

ACT-R must be implemented using the LISP programming language. The CMU ACT-R website offers numerous resources for helping potential ACT-R modelers in learning how to use the tool. Tutorials and manuals are available, free of charge, online http://act-r.psy.cmu.edu/actr6/. In addition, CMU hosts a summer school and workshops for learning to use ACT-R: http://act-r.psy.cmu.edu/workshops/ Further, CMU provides a contact person for support db30@andrew.cmu.edu

4.2.5 How Extensively Validated Is It?

ACT-R is extensively validated. Validation efforts are used to update and refine the cognitive theory behind the model.

4.3 Attention – Situation Awareness (A/SA)

4.3.1 What Is It?

The A/SA (Attention – Situation Awareness) modeling architecture was developed by researchers at the University of Illinois for NASA (Wickens, McCarley et al., 2008). It predicts operator situation awareness by using an attention model and a belief-updating module. The underlying attention model is SEEV (salience, effort, expectancy, and value).

SEEV predicts operator visual scanning, and says that scanning is driven by two bottom-up factors (salience and effort) and two top-down factors (expectancy and value). The salience, or obviousness, of a visual indication increases the likelihood that a cue will draw the operator's attention. Expectancy, the expectation that the information is changing frequently (events occurring rapidly) and therefore needs to be sampled often, also affects how often the operator looks at a display. Value, or the importance of information, increases the likelihood that the operator will view the information at the location of the valued commodity. Effort, the difficulty associated with moving attention to a display – either due to distance from the current visual focal point, or the need to navigate among pages in a multifunctional display – is the only factor that decreases the

likelihood of an operator viewing a display more often. Experienced operator scanning is driven relatively more by top-down factors (primarily by expectations and importance of information). The model runs as a discrete event simulation of SEEV, with probabilistic movement to attentional locations, to a degree proportional to the overall contributions of the four components.

Situation awareness is modeled according to two stages of the Endsley (1988) three-stage model, where stage 1 represents perception and stage 2 is comprehension (See Section 3.3 for elaboration of these concepts). For an operator to have awareness, s/he must perceive or attend to the relevant display. SEEV predicts whether or not the operator perceives the data. Once stage 1 awareness is attained, the belief-updating module in the A/SA architecture makes further predictions. A/SA models situation awareness on a scale of 0 to 1, where 1 is perfect awareness. Each display is associated with a particular type of awareness that the operator needs. When the display is viewed, awareness for the parameters on that display increases to 1. Over time, awareness decreases by a postulated decay function. For accurate information, the awareness decreases after a minute, but if distracting information is viewed, awareness decreases much more rapidly. The stage 1 SA of individual parameters is aggregated, and when sufficient parameter awareness exists, the operator is said to have achieved higher stages of SA. The overall awareness, in turn, drives scanning behavior. If the modeled operator has good SA, scanning is driven primarily by expectancy and value. Conversely, if the modeled operator has poor SA, sub-optimal scanning results. In the application reported in Wickens, McCarley et al. (2008), SA affects operator decision making, with better SA resulting in better decisions.

Visual scanning may be driven by habit, as in SEEV; but also by salient time stamped **events** that capture attention, such as the onset of a single warning indicator. The N-SEEV (Noticing-SEEV) model is an extension of SEEV (also developed by University of Illinois for NASA), and predicts operator noticing of a change in a visual field (McCarley et al., 2009; Sebok et al., 2006; Steelman-Allen & Wickens, 2011; Wickens, Hooey et al., 2009). The N-SEEV model uses SEEV to predict scanning, with the difference that a change occurs at some point in a simulation. The change is associated with a pre-defined location and salience in the visual field. Operator scanning continues, and if the operator views the changed display within a given time frame (typically within 10 seconds of the change), the operator is said to have noticed the change. If the operator does not view the change within that time frame, N-SEEV predicts that change blindness will occur, and the operator – even if s/he views the display – will not notice the difference. N-SEEV is typically run multiple times to generate a distribution of noticing times. The probability that the operator will notice the change is calculated based on the probability of runs that had fixations landing on the display where the relevant event occurred before the (e.g., 10 second) cutoff. The predicted time required for the operator to notice the change is calculated as the average noticing time for all of the "noticed" changes.

4.3.2 What Has It Been Applied To?

- SEEV: Aviation, Driving, Nursing
- A/SA: Aviation
- N-SEEV: Aviation

4.3.3 Where Can You Get It?

The A/SA, SEEV, and N-SEEV architectures are not software tools, nor are they particularly complex frameworks (like, e.g., ACT-R). They are practical models that can easily be implemented in a variety of programming languages. It is suggested that the models are developed with SME input, to identify the salience, effort, expectancy and value parameters for the visual displays in different operating conditions.

4.3.4 How Usable Is It?

The A/SA, SEEV, and N-SEEV architectures are easy to apply. No online support is available for these architectures but a wide variety of publications describes their implementation. The SEEV model can be implemented with simple analytic equations, although SMEs are required to estimate parameters of expectancy and value.

4.3.5 How Extensively Validated Is It?

A/SA, SEEV, and N-SEEV have all been validated against aviation human performance data.

4.4 The Cognitive Architecture for Safety Critical Task Simulation (CASCaS)

4.4.1 What Is It?

CASCaS (Cognitive Architecture for Safety Critical Task Simulation) was developed as part of a European Union project, HUMAN: Model-Based Analysis of Human Errors During Aircraft Cockpit System Design[1]. The intent for CASCaS was to address specifically the cognitive processes that are relevant for safety critical system design, particularly in aviation. As Lüdtke, Osterloh, and Frische (2012, p.1) state: "The innovative aspect of CASCaS is the prediction of human errors resulting from an interaction of (1) learned mental models (Routine Learning/Learned Carelessness), (2) actual limited cognitive performance and (3) safety nets in aircraft cockpit design (e.g. flashing indication, alerts and crew interaction)."

The HUMAN project included an effort to create an executable flight crew model that incorporated cognitive error producing mechanisms. This model was created using CASCaS, and it interacted with a flight simulator and a model of a cockpit interface. The model included routine learning, the learning of shortcuts, and attention allocation (Lüdtke, Osterloh, Mioch, Rister, & Looije, 2009).

The CASCaS model has been used to predict visual scanning behavior, dwell times in areas of interest (AOIs), and the time required to notice specific visual indications in the cruise and approach phases of flight (Lüdtke, Osterloh, & Frische, 2012). CASCaS, like ACT-R, is a cognitive architecture, which provides a structure and set of if-then rules for simulating human cognition.

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. These

[1] This project was performed during 1 March 2008 – 28 Feb 2011. More information is available at:
EU website: http://ec.europa.eu/research/transport/projects/items/human_en.htm
HUMAN Aero site: http://www.human.aero/

correspond, respectively, to the skill-based, rule-based, and knowledge-based levels of behavior defined by Rasmussen (1983). New tasks require significant effort on the part of an operator, and are performed at the cognitive level. These are modeled by the operator having a high-level goal, and selection of a plan by which to meet that goal. When operators are somewhat familiar with a task, they are working on the associative level. These types of tasks are modeled by selection of a set of rules. Finally, frequently performed tasks are autonomous, and performed without conscious thought. These include maneuvering the aircraft (Lüdtke, Osterloh, Mioch, Rister, & Looije, 2009).

In solving problems, humans typically apply the easiest solution – autonomous, if possible; associative, if they can identify a relevant pattern for the situation; and cognitive if absolutely necessary due to lack of established rules or safety-critical concerns. The Lüdtke, Osterloh, et al., (2009) team focus on error mechanisms at the associative and cognitive levels.

Knowledge in the associative and cognitive layers is stored in the memory component. Short term memory includes data that have been perceived by the modeled operator from the environment or derived from rules. Long term memory stores flight procedures as "if then" rules. When the "if" condition is met, the "then" condition is triggered. Figure 4.1 below shows an overview of the CASCaS components.

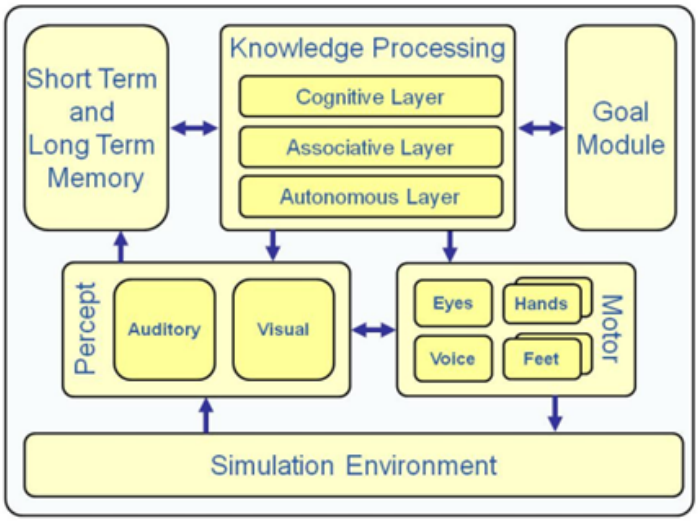

Figure 4.1. Overview of CASCaS Components
From http://www.isi-padas.eu/sites/default/files/ISi-PADAS_Newsletter_5.pdf

In another CASCaS effort (Lüdtke, Osterloh, Mioch, Rister, & Looije, 2009), the CASCaS model predicts cognitive errors such as Learned Carelessness and Cognitive Lockup (See 3.1 and 3.4 for descriptions of these applications).

4.4.2 What Has It Been Applied To?

CASCaS has been applied to a variety of aviation tasks, but primarily focusing on pilot error mechanisms. It has also been applied to modeling vehicle driver performance (http://www.isi-padas.eu/sites/default/files/ISi-PADAS_Newsletter_5.pdf).

4.4.3 Where Can You Get It?

It does not appear that CASCaS is available, neither as open source product nor for purchase. For more information, the reader should contact Dr. Andreas Lüdtke (Lüdtke@offis.de).

4.4.4 How Usable Is It?

This is not relevant, since the architecture is not available.

4.4.5 How Extensively Validated Is It?

CASCaS has been quite extensively validated as part of the EU project that funded it. The effort included many modeling phases as well as empirical, pilot in the loop studies. Validation efforts focused on particular aspects of performance (e.g., when an error occurred in seeking additional, safety-critical information).

4.5 The Man-machine Integration Design and Analysis System (MIDAS)

4.5.1 What Is It?

The Man-machine Integration Design and Analysis System (MIDAS; http://humansystems.arc.nasa.gov/groups/midas/) is an established HPM that predicts human-system performance under nominal and off-nominal conditions (Gore, 2010). MIDAS is a dynamic, integrated human performance model environment that facilitates the design, visualization, and computational evaluation of complex man-machine system concepts in simulated operational environments. MIDAS symbolically represents many mechanisms that underlie and cause human behavior including the manner that the operator receives/detects information from an environment, comprehends and registers this information in a memory store, decides on a response, and responds to the information within the context of operational rules and human performance capacities. MIDAS combines these symbolic representations of cognition with graphical equipment prototyping, dynamic simulation, and procedures/tasks to support quantitative predictions of human system effectiveness, and improve the design of crew stations and their associated operating procedures. MIDAS provides an easy to use and cost-effective means to conduct model simulation experiments that explore "what-if" questions about domains of interest. Figure 4.2 illustrates the organization and flow of information among the MIDAS components. For a description of the MIDAS processes, the reader is directed to Gore (2010).

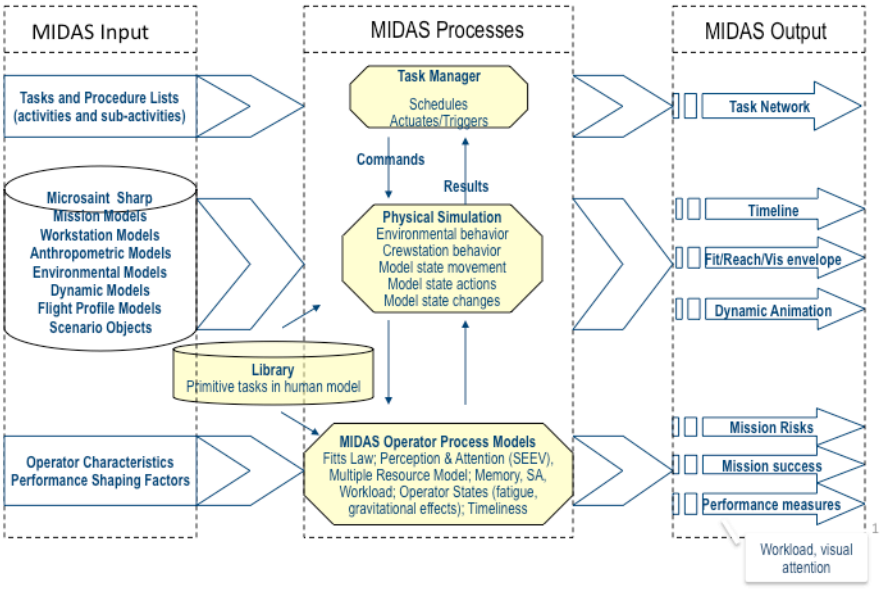

Figure 4.2: Illustration of the flow of information into MIDAS v5.

MIDAS gives users the ability to model the functional and physical aspects of the operator, the system, and the environment, and to bring these models together in an interactive, event-filled simulation for quantitative and visual analysis (Hart et al., 2001). Operator behavior within a MIDAS simulation is driven by a set of user inputs specifying operator goals, procedures for achieving those goals, and declarative knowledge appropriate to a given simulation. These asserted knowledge structures interact with and are moderated by embedded models of perception for extracting information from the modeled world and embedded models of cognition for managing resources, memory, and actions. In this way, MIDAS seeks to capture the perceptual-cognitive cycle of real world operators who work towards their most immediate goals given their perception of the situation and within the limitations of their physical and cognitive capacities. In MIDAS, as in the real world, perceived changes in the world - new information or events - may cause changes in the adjudged context of the situation triggering new goals or altering methods to achieve current goals. Such perceived changes may be precipitated through the behavior of other modeled entities or by user specified time-based, condition-based, or probabilistic events set within the simulation (e.g., a system failure or the receipt of an incoming air traffic control clearance). It may, in fact, be the impact of the operator's own actions which lead to changes in context and goals.

In sum, MIDAS v5 can be used to: develop guidelines (e.g. for NextGen aviation roles and responsibilities and information requirements), design and test procedures, conduct "what-if" scenario evaluations, design and evaluate cockpit designs and information placement, evaluate multi-crew performance designs, predict situation awareness profiles of multiple operators, generate stochastic human performance, predict attention drivers, predict cockpit-related attention overload, generate human operator workload timelines, and determine optimal strategies for human-automation interaction.

Inputs: Data from psychology experiments for the empirically-driven operator models (human perception, cognition, memory, decision, and behavioral models), task network models input into a discrete event simulation tool to represent the tasks and activities required of the human in specific contexts, visualizations of anthropometric characters and of the flight deck and other crewstation

assumptions (e.g., using a computer aided design [CAD] model for a Boeing 747 cockpit), display / environmental information context weights (to populate the attention and situation awareness models).

Outputs:
Overt/observable behavior, time to perform tasks, task accuracy, and task completion / success / failure, human error given task failure, operator workload (visual, auditory, cognitive spatial, cognitive verbal, fine motor, gross motor, and vocal), task/environmental timelines, current ongoing tasks, situation awareness profiles, the task network of the required operator actions, CAD of candidate cockpit designs, operator strategies for human automation interaction and integration.

4.5.2 What Has It Been Applied To?

Primarily to civilian cockpit design, but applications of MIDAS have also been made to space systems (crew vehicle exploration, ISS interface, robotic arm control), military helicopters, and air traffic control. The full list of MIDAS application domains can be found online at http://humansystems.arc.nasa.gov/groups/midas/applications.html while the publications reporting on these and on the empirically driven operator models can be found online at http://humansystems.arc.nasa.gov/groups/midas/publications.html

4.5.3 Where Can You Get It?

A software usage agreement with NASA Ames needs to be obtained to use the software (MIDAS NASA Civil Servant POC: David Foyle – David.C.Foyle@nasa.gov). The MIDAS software operates on the Windows NT platform and requires the freely available .net framework available from the Microsoft website.

4.5.4 How Usable Is It?

MIDAS has proven to be a flexible and useful tool for representing humans operating in a variety of environments. Its greatest strength lies in its ability to represent conceptual designs or candidate procedures in software and then allow a designer or potential user see them operated by a virtual crewmember in the context of a simulation of the target environment. The graphical user interface, timeline output, and anthropometric visualization have made MIDAS accessible to a broader range of potential users. The code has been used extensively over the previous four years. However, from the perspective of a formal usability evaluation determination, given that the MIDAS v5 user guide is currently under development, it is best to consider MIDAS an internal R&D software tool. The software can be downloaded onto a Windows NT platform from a website.

4.5.5 How Extensively Validated Is It?

MIDAS has undergone extensive verification and validation efforts. Each of the component models of human behavior has been validated through empirical human simulation comparison. The component models were then computationalized (represented in algorithmic form) and verified within the MIDAS architecture through extensive verification exercises. The integrated architecture (which contains the host of validated component models) has been validated in (Gore Hooey, Socash et al., 2011).

4.6 Multiple Resource Model

4.6.1 What Is It?

The multiple resource model predicts the success or failure of multi-tasking in the cockpit. The architecture of the multiple resource model is generally built around the multiple resource theory of multi-task performance (Wickens, 1984; 1990; 2005; Wickens, Bagnall, Gosakan, & Walters, 2011, North & Riley, 1989). It is an analytic model that contains two macro elements.

- A **conflict matrix** that predicts the effect of competition for specific (multiple) resources between two time-shared tasks.
- A **demand component** that predicts the interfering effect of the difficulty of the two tasks.

Each is described in turn.

The conflict matrix. The multiple resource model (Wickens, 2008b) assumes that human information processing depends upon some or all of levels along four dichotomous dimensions, as shown in the table. Table 4.2.

Table 4.2. Dimensions of the multiple resource model (modalities now also includes tactile in the latest version)

DIMENSION	**Level 1**	**Level 2**
Stages	Perceptual-cognitive	Action selection
Codes	Verbal/linguistic	Spatial
Modalities	Visual	Auditory
Visual channels	Focal	Ambient.

Note that there is some nesting within these. For example the two visual channels are nested only within the visual modalities, and the different modality channels are nested only within perception/cognition. Note also that the latest version of the multiple resource model now contains Tactile, as a 3rd level of Modalities (Wickens, Hollands et al., 2012)

Any given task can be identified as occupying one or more cells of the matrix. For example talking to ATC will occupy action selection and verbal/linguistic. Thus any pair of tasks may vary in the extent to which they occupy the same cells or different cells. The key to computing resource conflict, is that two tasks will interfere with each other to the extent that they occupy more overlapping cells. The computation mechanism of such prediction is based on the conflict matrix between the tasks. To simplify, if only two resource dimensions were specified (stages and codes) such a conflict matrix would be shown as in Table 4.3

Table 4.3. Representation of the conflict matrix in a simplified four-resource version of the multiple resource model.

Task A / Task B	Perceptual-cognitive spatial	Perceptual cognitive verbal	Action Selection - spatial	Action Selection - verbal
Perceptual – cognitive spatial	.XX	.YY	.ZZ	.AA
Perceptual – cognitive verbal		-	-	-
Action Selection - spatial			-	-
Action Selection - verbal				-

Then the entries within each cell will show how much conflict there will be within a cell demanded by both the task labeled across the rows, and that labeled down the columns. For example if Task A is responding to ATC (response verbal) and Task B is examining a map for route vectors (perceptual - cognitive spatial) the conflict interference between them will be 0.AA. These interference values are fractions ranging from 1 to 1.0. Details on specific numbers can be found in Wickens, 2002a; 2005). Thus any pair of tasks, whose resource demands are ticked off along across the top, and down the sides, will populate a certain number of cells within the matrix. The conflict values for those populated cells are summed to achieve a total resource conflict score.

Total Demand score

Table 4.4 now depicts the original task outline from Table 4.3, within the framework of an outer row and column. These explicitly label the two tasks (here map reading and target searching) and assign a demand vector to each: map reading is heavily perceptual/cognitive and spatial, but here may involve a simple voice response. Visual search (e.g., traffic search out the window) demands only visual perception. The total task demand is simply the sum of the demands of the two tasks, here 2 + [3+1] = 4. Also in this example it can be seen that the conflict score is [.XX + .AA] (the sum of the two populated cells). Hence the total interference is predicted to be the weighted sum of the demand score and the conflict score.

Table 4.4 The simplified four-resource model depicting the task demand vectors of the two competing tasks, map reading and target search.

	Map read	**3**	**0**	**0**	**1**
Target search		Perceptual-cognitive spatial	Perceptual cognitive verbal	Action Selection - spatial	Action Selection - verbal
2	Perceptual – cognitive spatial	**.XX**	.YY	.ZZ.	**.AA**
0	Perceptual – cognitive verbal		-	-	-
0	Action Selection - spatial			-	-
0	Action Selection - verbal				-

Note that in this computational architecture, the two components are treated independently and do not interact. That is, the ".XX" value in the upper left had cell, will be the same with the demand vectors of 2 and 3 (as shown) as it would be if these were, for example, 1 and 1 (two very easy perceptual tasks). Riley et al. (1991) have provided data to show that this simple non-interacting version is fully adequate.

Within the model, the conflict values within the matrix are relatively stable and fixed within the model. However a task analysis is required by the model user to establish which resources are demanded by each task, and the degree of demand within each resource. The latter may be estimated by SMEs, or in some cases, may be available from table lookups, such as the McCracken and Aldrich scale.

The depicted model is analytic, and its computation is straightforward. However two elaborations can be made to provide a discrete event simulation model:

1. As noted in the TLAP architecture (See 4.7 below), if task times are variable, then the overlap between two concurrent tasks will vary according to task time distributions. Then multiple resource conflict and demand should only be computed during the periods when the two tasks overlap in time. During epochs of single task performance, multiple task interference will not (by definition) be occurring.

2. This version of the model assumes a level of interference as computed above. But in reality, if a pilot is confronted with two tasks, whose demand, or resource conflict is so high that it is impossible to process them concurrently, then, by necessity, she must shed one or the other. This task shedding behavior is based on a red-line assumption which is not itself an inherent part of the multiple resource model, and is addressed instead in the architecture of scheduling models such as the scheduling module for task management within MIDAS (Gore, 2013, in preparation; see Section 4.5).

4.6.2 What Has It Been Applied To?

Primarily to flying applications as described in 3.2; although there is one application to driving with in-vehicle technology (Horrey & Wickens, 2004).

4.6.3 Where Can You Get It?

The multiple resource model is not available as a stand-alone software package. However versions of this model have been incorporated into MIDAS (see 4.3 above), and into IMPRINT. IMPRINT is available (upon request) to U.S. Government agencies at the following URL: http://www.arl.army.mil/www/default.cfm?page=445.

4.6.4 How Usable Is It?

Approximations to computing the two components (conflict and demand) are relatively easy to accomplish. Demand values can be assigned by SME's. The demand vectors associated with each task can also be assigned by SME's. Different sets of conflict values within the conflict matrix are available in the published literature (e.g., Wickens, Bagnall et al., 2011).

4.6.5 How Extensively Validated Is It?

Several flight deck validations of the computational model were described above. A single driving validation is described in Horrey and Wickens, (2004). Other less quantitative validations of multiple resource predictions in a flight deck environment have been carried out by Wickens, Sandry, and Vidulich (1983; military cockpit), and Wickens, Goh et al (2003; GA cockpit with NextGen technology).

4.7 Optimal Control Model (OCM) & Flight Control Workload

4.7.1 What Is It?

The Optimal Control Model (OCM; Rikard & Levison, 1981) is actually a prototype of a more general class of manual control models that have been employed (and validated) to predict flight handling qualities (see Wickens, 1986); and the latter commodity is closely related to subjective workload, or more specifically psychomotor load (within a VACP context).

The general architecture of OCM (Kleinman, Baron, & Levison, 1971; Levison, 1989) is to model three fundamental stages of pilot information processing on the flight deck. These are:

1. ***Estimating*** the state of the vector of displayed variables necessary to control an axis of flight (e.g., vertical flight, lateral flight). An optimal Kalman filter is incorporated to do this estimation.

2. ***Predicting*** the future level of these variables to the extent they are subject to lag, and that stable control depends on predicted state, rather than current state.

3. ***Translating*** the estimated predicted state into a vector of control gains, applied to different flight controls. This gain matrix is tuned to *optimize* the balance between the need for precise control (minimizing error) and "smooth flight" (e.g., minimizing large rapid control movements).

All three stages contain optimization components, hence the name "optimal" control.

Importantly, these processes are subject to, and hence modeled with:

- Inherent lags in the human processing system (default, around 1/3 second)
- Noise added to the estimation and control processes.

The most important noise is that added to estimation and observation, (called **equivalent observation noise**) and is assumed to grow linearly as **visual attention is allocated away from the controlled variable in question** (Levison, Elkind, & Ward, 1971), This function provides an important mechanism for multi-task performance and workload prediction.

4.7.2 What Has It Been Applied To?

The model has been primarily applied to flight handling qualities, although one application has been to driving (Levison, 1989).

4.7.3 Where Can You Get It?

The model was at one time available from Bolt, Beranick, and Newman. It is not clear if it still is available.

4.7.4 How Usable Is It?

The OCM is based on the frequency domain language of linear and non-linear feedback control theory. Hence it is designed to work on analog computers (or digital simulation of linear dynamic systems in the frequency domain). As a consequence, its use requires some degree of specialized knowledge in control theory, generally taught within the discipline of aeronautical engineering.

4.7.5 How Extensively Validated Is It?

The model has been validated in a variety of applications to predict flight control performance and workload. The Rickard and Levison (1981) study cited is prototypical of a number of others reported in various proceedings of the Annual Conference of Manual Control.

4.7.6 Extensions of OCM

Other models of manual control within the frequency domain have also been produced and validated to predict flight path tracking performance and occasionally workload. Perhaps the most common of these is the **Crossover model** of McRuer and Jex (1967). We do not review this model here because it is not really a model of human performance, but rather, a model of the whole human-aircraft control system. However we do note that one important feature of this model (and other models of manual control), is the close correlation found between model-predicted lag of an aircraft axis, and measures of handling qualities (assessed by the Cooper-Harper rating scale) which we noted above, is a proxy for the subjective estimate of psychomotor workload. It is assumed that when there is lag along flight control axes, the pilot must “generate lead” (e.g, predict) in order to compensate for the lag and retain flight stability. Prediction is workload-demanding (Wickens, Gempler, & Morphew, 2000) and the greater the need for prediction in flight control, the greater the workload (McRuer & Jex, 1967; Wickens, 1986; figure 39.31).

4.8 The Traffic Organization and Perturbation Analyzer (TOPAZ)

4.8.1 What Is It?

The Traffic Organization and Perturbation Analyzer (TOPAZ) is a multi-agent dynamic risk model designed to evaluate system safety within Air Traffic Management (ATM). The system uses a dynamic extension of Petri nets and Monte Carlo simulations to analyze the safety of air traffic systems through the generation of conditional collision risks. TOPAZ can account for both nominal and non-nominal events and can represent the dynamic interactions between human operators, technical systems and procedures.

TOPAZ has been under development for a number of years. The initial system included a simulation environment based on high-level petri nets. This included support for developing petri net modules for human behavior, the environment and systems within the ATM environment. Additions have included multi-agent situational awareness modeling, support for bias and uncertainty assessment, and modules for accident model specification (Stroeve et al., 2009).

The system supports the development of *systemic accident models* describing the performance of the system as a whole in which accidents become emergent properties of the variability within the system (Stroeve et al., 2009). This quality is important for NextGen in that it supports the evaluation of very low frequency events, such as runway incursions, for which little accident data exists.

The process of developing the model involves representing the dynamic characteristics of related agents, such as human operators and technical systems, within a hierarchical structure. These model structures can represent:

- Key aspects of agents including SA, task scheduling, flight phases, alerting systems
- Agent modes related to human and system performance
- Time characteristics of tasks or systems such as event distributions
- Agent interactions such as task transitions or system availability
- Interactions including the effects of SA on performance and the effect of system availability

(Stroeve et al., 2009)

Numerous Discrete Event Simulation Monte Carlo runs of the model then provide conditional probabilities of incident occurrence based on the dynamic interactions of agents and interactions represented. The TOPAZ system also supports both bias and uncertainty analysis allowing the user to assess the relative contribution of agents and events to the overall system risk.

4.8.2 What Has It Been Applied To?

The primary use of TOPAZ has been the evaluation of collision probability within the ATM environment. As mentioned in the Error (3.1) and SA (3.3) sections, the most common scenario evaluated has been runway incursions resulting from one taxiing aircraft erroneously crossing a runway in which another aircraft is on a take-off roll. The model includes dynamic representations and interactions of the following agents:

- The motions of the take-off and taxiing aircraft through the take-off roll and approach to runway intersection respectively
- The availability and dynamics of the surveillance system including relative position and velocities of each aircraft and the resulting alerts and alarms
- The pilots flying the two aircraft and a range of performance dynamics such as visual monitoring, conflict detection and reaction
- The runway and taxiway controllers and performance dynamics such as communication, conflict detection and reaction.
- The availability and dynamics of the communications systems between the controllers and pilots including delays, frequency selection and the nominal state of these systems.

Figure 4.3 shows the relationships between the agents. The scenario begins with the take-off roll and taxiing of the two aircraft (nodes E10 & E12) and ends with the probability of detection and avoidance or the collision. In one analysis, a huge number of Monte Carlo runs were used to generate the probability of the collision event. This included sensitivity analysis removing agents or varying aspects of agent behavior to assess the effects on the overall risk.

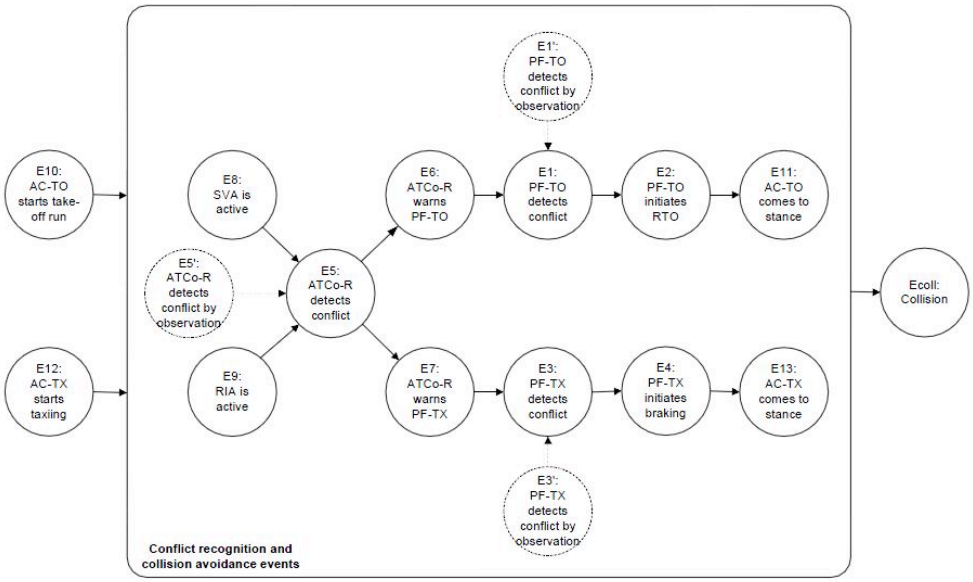

Figure 4.3. Shows the relations events for conflict recognition and collision avoidance actions by the agents in the runway incursion model (from Stroeve et al., 2011).

The results of the bias and uncertainty analysis show that the understanding of pilot performance is the strongest contributor to uncertainty in the risk results. This result would help to direct the analysis team to focus data gathering and modeling efforts on ensuring the accurate understanding and representation of pilot performance within the model. In addition, an analysis of the conditional probabilities of each event given a collision illustrates the individual contributions of each agent to the overall risk. For example, overall collision risk only increases by a factor of 1.06 without an ATC alert system but increases by a factor of 500 if none of the agents (pilots and controllers) are monitoring or in the control loop (Stroeve et al., 2011).

4.8.3 Where Can You Get It?

TOPAZ has been developed by and is available from the National Aerospace Laboratory NLR, Air Transport Safety Institute, Amsterdam, The Netherlands www.nlr.nl.

More information about the model can also be obtained by contacting Henk Blom. (Henk.Blom@nlr-atsi.nl)

4.8.4 How Usable Is It?

TOPAZ is a complex system requiring considerable expertise and effort. The process includes identification of the operation and hazards to be modeled and the development of high-level Petri nets for the various agent models.

4.8.5 How Extensively Validated Is It?

The studies describe a mathematical process of verification rather than validation. Actual data with which to compare model outputs in a true validation are difficult to obtain for such low-probability events.

4.9 Time Line Analysis Procedure (TLAP)

4.9.1 What Is It?

This model has a simple straightforward architecture (Parks & Boucek, 1989). A time line is laid out concerning when different cockpit tasks need to be performed. Each task is assigned a length, which may be fixed (in the analytical version) or have a variance (in a discrete event simulation version). Tasks may overlap in time (as flying the aircraft, while communicating with ATC). Then the total time of a mission is subdivided into time units (e.g., 10 seconds), and within each unit, workload is computed as the ratio of the total of task times within that unit, to the unit length. For example if there are two 5 second tasks within the 10 second unit, the workload is 10/10 = 1.0 (or 100%).

Note that at this simple level, the model does not distinguish whether the two 5 second tasks are performed sequentially or simultaneously. A more refined version of the model adds a penalty for two tasks occupying the pilot at the same time (e.g., simultaneously) and this version has been found to be a more accurate predictor of performance breakdowns (Sarno & Wickens, 1995).

The model can be converted from an analytic model (calculating ratios) to a discrete event simulation model, if the time required by each task has a variance associated with it, and each iteration picks (randomly) a time demand from the distribution for each task. The time line analysis procedure does not incorporate specific assumptions about the pilot response when there is concurrence (e.g., possible shedding or delaying the task, as in a queuing theory model), nor does it consider the nature of the task in question, other than its length. These issues are addressed in the Multiple Resource Architecture.

4.9.2 What has it been Applied To?

As described in section 3.2, it has been applied to many cockpit time-line procedures, with Stone et al. (1987) and Parks & Boucek (1989) offering prototypical examples.

4.9.3 Where Can You Get It?

The original Time Line Analysis (TLA) module was part of a software tool called Crewstation Human Engineering Software System (CHESS), it was used for the Boeing 757/767/747-400 certification. It has not been employed subsequently for certification, and does not currently exist within a commercially available software tool.

4.9.4 How Usable is it?

TLAP is extremely usable, and any user can follow the general guidelines presented in section 3.2.4; or outlined in Parks & Boucek (1989) to create his/her own analysis.

4.9.5 How Extensively Validated is it?

In our review, there are surprisingly few high validity (e.g., commercial flight deck) validations against performance data or task shedding data; given the ease of use. But see Sarno & Wickens (1995) for lower validity validation.

4.10 GOMS

4.10.1 What Is It?

The GOMS (Goals Operators Methods, Selection rules) is an approach to task analysis that is based upon the keystroke level model of human information processing originally crafted by Card, Moran, and Newell (1983). This approach provides fundamental times for basic operations, like looking, item rehearsal, or reaching (as a function of distance and target width). Following a GOMS task analysis, the analyst can assemble tasks so as to create a network of more complex operations, necessary to complete higher-level tasks, like programming an FMS (Polson, Irving, & Irving, 1994; John et al., 2009). The GOMS approach has been used together with the ACT-R modeling architecture to create human performance models that provide predictions for both cognitive errors (ACT-R) and time to complete task sequences (GOMS).

Gil et al., (2012, in press) have enhanced the basic GOMS tool, described in Polson and Javaux (2001), to create E-GOMSL (Enhanced GOMS language), an architecture that is more versatile and appears to be more specifically focused on cockpit tasks. E-GOMSL identifies a set of basic **operators:** [confirm, think-of, look-at, store, recall]. Each operator in turn is characterized by a set of four attributes: [Channel, VACP; Control object (e.g., working memory), syntax, and time (or time distribution)].

Considerable effort in task analysis appears to be required to assemble the operators to approximate the cockpit tasks; but the model will output task times as well as variables such as working memory load, that are important for predicting workload (Gil & Kaber, 2012).

The Model, and the requirements for programming, is well described in Gil and Kaber (2012).

4.10.2 What Has It Been Applied To?

The GOMS approach to task analysis has been applied extensively in analyses of human-computer interaction. It could potentially be used to model any detailed level of interaction in which an agent performs discrete action sequences. There have been some variants of GOMS applied in the aviation environment. These variants include the Cognitive, Performance, Motor (CPM-GOMS). CPM GOMS has focused on F22 operations as well as predicting mouse clicks for controllers

(Remington, Matessa, Freed, & Lee, 2003). A more recent application of the GOMS approach that combines two theory-based tools (CogTool and SANLab) has been used to address human variability in skilled performance as applied to the Boeing 777 Flight Management Computer (FMC) and the Control and Display Unit (CDU) (John, Patton, Gray, & Morrison, 2012). This paper contains no validation data.

4.10.3 Where Can You Get It?

The GOMS method can be applied to any tasks that involve keystroke-level type interactions. Information on this method is available in a variety of sources, including:

Kirwan, B. & Ainsworth, L.K. (1992). *A Guide to Task Analysis*. Washington, DC: Taylor and Francis.

Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human–Computer Interaction*. Hillsdale, NJ: Erlbaum.

http://en.wikipedia.org/wiki/GOMS

4.10.3 How Usable Is It?

GOMS is a potentially time-consuming method for analyzing task sequences, but it can provide useful insights for detailed task sequences. The only usability concerns are related to learning the method itself; there is no particular software that the analyst needs to learn or use.

4.10.4 How Extensively Validated Is It?

GOMS has been in use for 3 decades (as of the date this report was produced) and numerous validations have been conducted using the method (e.g., Gray, John & Atwood, 1993).

5. DISCUSSION and CONCLUSIONS

5.1 Validation Efforts

In evaluating the overall success of the aviation modeling community in generating valid pilot performance models, our findings were somewhat disappointing. In Table 2.4 (see Section 2.2.4), we reported an overall validation rate (across all model aspects) of 36%; and many of the higher values contributing to that overall percentage were aspects not reviewed in our deep dives (e.g., pilot vision models, manual control). Of the six deep dive areas, where we paid close attention to the quality of validation, the percentage was smaller, as shown in Table 5.1, which breaks down validation by whether or not it contained a quantitative aspect.

Table 5.1 Validation statistics for deep dive aspects

Model aspect	All validations	Included Quantitative aspects
Error	9/17	0/17
Workload/multi-task	14/33	11/33
Situation awareness	3/15	1/15
Pilot-automation interaction	8/16	3/16
Roles & responsibilities	3/12	2/12
TOTAL	37/93= 40%	17/93 = 18%

Several reasons can be offered to account for this state of affairs. First, and of relatively minor importance, the statistics reported could be adjusted upward or downward slightly depending on how validation is defined and classified for each model effort. For example some of what we said was ***qualitative*** validation (and hence counted in the left column but not the right) might be argued to be quantitative (e.g., if **some** numerical data were provided from a PITL simulation, but it was challenging to translate these 1-for-1 into the model predictions, e.g., workload was assessed by NASA TLX in an empirical study, yet calculated as VACP scores in the model).

Furthermore, some very good model **verification**, (and we emphasize that a much higher percentage the model architectures reviewed *were* verified, as shown in Table 3.6.1) could be argued by some observers to be a form of ***qualitative validation*** (hence boosting the numbers in the left column of Table 5.1). But at the same time, we might have classified a model as having a qualitative validation, and other observers would challenge this as being a verification, given the fuzziness of the boundaries between these two concepts. Nevertheless we would argue that the precise values of the statistics reported are less important than their approximate level; which indicates that well less than half of the model efforts received validation, and less than half of those received the sort of precise numerical quantitative validation that can both "sell" the model as being accurate (if the correlation is high) and provide precise guidance for model modification, if the correlation is lower. We do observe however in Table 4.1, that all the major architectures have received considerable validation. On this basis, many of the predictions from such architecture made in *model efforts* that have not yet been validated, are still extremely useful because the prior validation of the architecture, provides confidence in the accuracy of the predictions.

To elaborate on the concern about the state of validation, we note that the numerators of several of these statistics in Table 5.1 are heavily populated by studies involving non-transport aircraft and/or non-transport pilots as apparatus or subjects. And even of those studies that used professional pilots as participants, using aircraft simulations representing modern commercial carriers, only a small handful directly examined NextGen issues to test their models predictions.

We offer four non-mutually exclusive reasons for the state of affairs regarding validation.

First, finances for the researcher/modeler are often limited. Most models are developed under contract; and the time and effort to accomplish the full validation at the end of the contract are often underestimated, following the well-known *planning fallacy* (Buehler, Griffin, & Ross, 2002). Funding runs out before complete validation can be accomplished. Furthermore we have noted that in at least two cases, a model was developed in phase 1 of a 3-phase SBIR, and subsequent phases (in which validation was planned) were not funded.

Similarly, along the lines of "limited funding," models provide the opportunity to collect a tremendous variety of data regarding predicted pilot performance. Empirical validation analyses, due to time or resource constraints, sometimes focus on a limited set of the data. Thus sometimes the validation is carried out on only a small sampling of the total data set, such as one set of subjective ratings or one aspect of visual scanning.

Second, validation is cumbersome; particularly in high fidelity PITL simulations. Pilots and simulator time are hard to obtain. Furthermore even when simulations are accomplished, certain kinds of data are hard to quantify, such as "a pattern of errors" (see Section 3.1), or a particular sequence of procedural steps. Also careful experimental design is necessary to create the **different**

experimental conditions necessary to accomplish true validation. That is, is the model sensitive to **changes** in flight procedures or equipment in the same way that pilots are. It is only by comparing model-predicted and pilot-predicted **differences** or **changes** that these critical correlational measures of validation can be obtained.

Third, some validation studies appear to report narrowly focused, promising results rather than a complete set of results. The more comprehensive pilot models (e.g., those created using MIDAS, CASCaS, or ACT-R) can predict many aspects of pilot behavior (e.g., errors, workload, situation awareness), yet several of the validation results provided only data on highly specific issues (e.g., visual workload, errors in selecting a particular page of data). This suggests that the researchers might be reporting only those that indicate good agreement with the model. Alternatively, it might indicate a bias in the literature, where "uninteresting" results (e.g., of a low correlation prediction) are typically not published.

Finally, sometimes numbers may be available from the PITL data, but for reasons that are unclear to the authors of this report, were not converted to appropriate statistics. In this regard we wish to reiterate that the most appropriate statistics are measures of the shared variance accounted for by model and PITL simulation results, where this variance is across-conditions that are meaningful to the FAA (e.g., old vs. new procedures, performance with and without NextGen technology or the 4-way combination of these two). This shared variance is of course represented by the correlation, with N defining the number of conditions compared (see 2.2.3 and Appendix A).

However in this regard we also want to emphasize a point made in Section 2.2, that it is the **raw value** of the correlation (from 0 to 1.0) that is important, not necessarily its statistical significance. This is of course because the latter is influenced by a sample size, and if only 3 or 4 conditions are available for model validation, imposing constraints on statistical significance on a 4-point correlation is an impossibly high bar. Indeed even when only two conditions are available (e.g., conventional vs. NextGen cockpit), and hence a correlation is meaningless, useful quantitative validation can be achieved by comparing the percent *difference* observed in pilot performance (e.g., a 30% reduction in errors) with that predicted by the model. This corresponds to the slope of a regression line, which will be closer to 1.0 as the model is more valid.

5.2. Status of Flight Deck Models

Notwithstanding the limitations of the validation efforts, our team has identified an impressive array of models that either are, or could be tailored to address, issues in NextGen. These models were defined by those "aspects" of pilot performance that they predict, and these aspects could in turn be sorted into three different categories, ordered perhaps from least to most relevant.

First, there were aspects such as spatial disorientation, decision making and manual control that are relevant to all aspects of aviation, and for which there was not time nor resources to pursue in depth. It was decided during our mid-contract meeting with FAA sponsors that these would need to await further study because of our own limited resources.

Second, there were two particular model aspects that we did not initially intend to deep dive, but as we proceeded into the second half of our contract effort we realized were of great relevance to the six aspects that we were pursuing in depth. These were **procedures** and **vision** (particularly **visual attention**). Indeed both of these were partially reviewed in the deep dive effort of Section 3, because

of their high degree of relevance: for example procedural models, such as GOMS are often relevant to activities like programming an FMS, covered in our PAI aspect. Correspondingly, models of visual attention are sometimes a necessary component of situation awareness and automation (PAI) modeling. We believe that these two areas should receive high priority for future evaluation and model development. We also note that vision models in particular were identified early on in our project to have benefitted from a good deal of validation efforts (63% in Table 2.4), and some of this validation was in fact reported in Section 3.

Third of course, are the six model aspects that did receive our deep dive analysis (there were five categories in Section 3, because workload and multi-tasking were combined into one section). After considering these models in some depth, it is our conclusion that, ideally, and at this point in the development of NextGen, they should not be pursued in isolation but rather that architectures should combine aspects in a manner similar to that typified by, for example CASCaS, MIDAS and TOPAZ. Such combination is warranted because the pilots' responsibilities in NextGen are complex and, themselves, integrate these aspects. Several examples abound:

- The pilot using automation (**PAI** aspect) must maintain **situation awareness** of automation state often through **visual attention**
- The pilot's task of programming the **procedures** of complex automation (**PAI**) imposes heavy **workload**
- Different assignment of **roles and responsibilities** can substantially influence **multi-tasking** requirements, as well as impose and/or influence **communications load**.

Each of these three examples link three aspect areas, and for their impact to be predicted, the relevant model should, ideally, incorporate the set of relevant aspects.

Of course with the increasing complexity of multi-aspect models comes the danger of decreased usability for teams and individuals other than the model developer; although it was not our goal in the current effort to "score" models along such a usability scale. Still, we would argue that developers of complex models be sensitive to these issues, and perhaps consider a modular "plug in" concept, whereby a given model aspect, perhaps vital to answer some modeling questions, can have a "default' (and simplified) representation if that aspect is not critical for a particular application. For example a model might have a simplified representation of workload, not requiring channel and resource assignment that could be employed when workload questions are not critical; but could be replaced by a more sophisticated workload model when they are critical

5.3. Final Conclusions

The following list summarizes our main conclusions from this review and analysis:

- There are many modeling architectures available for predicting pilot performance.
- Numerous efforts have been conducted to predict pilot performance.
- There appears to exist a direct relationship between breadth and complexity of modeling efforts (a model that predicts many aspects of pilot performance will probably require extensive modeling expertise).

- Validation efforts are currently insufficient, particularly in model aspects of communications, procedures, and roles & responsibilities, all of which have less than around 25% validation rates.
- Validation efforts should be accompanied by explicit correlation measures, and such measures accompanied by the raw data scatter plots from which they are derived. In this way readers and researchers can inspect the particular conditions (model predictors) that may be "outliers", either under- or over-predicted by the model.
- Verification efforts differ in the extent to which they have been implemented.
- There is a need for clear standards in terms of how models are verified and validated.
- Development of models that address multiple aspects in combination should be encouraged.
- There is a need for more funding for model verification and, especially, validation efforts.

Based on these findings, we recommend the following components for a "highly rated modeling effort": that it include an empirical validation component, that it be scoped appropriately (focused if possible), and that it use an appropriate "tool for the task": one of the broader modeling architectures if it is attempting to model a diverse range of behaviors (e.g., MIDAS, CASCaS, TOPAZ); a cognitive model for addressing cognitive issues (e.g., ACT-R), that it be reasonably well validated. Such an effort should include a clear specification of different NextGen procedures or technologies; and every effort should be made to maintain features of the PITL simulation identical to the conditions of the model simulation. Correlation data should be reported in scatter plot form

We would like to conclude that modeling efforts contribute significantly to improving aviation safety, by providing a viable and useful test bed for evaluating design decisions well before implementation. Modeling efforts provide insights into human cognition and behavior, and offer useful input to system designers

6. REFERENCES

Aldrich, T. B., Szabo, S. M., & Bierbaum, C. R. (1989). The development and application of models to predict operator workload during system design. *Applications of human performance models to system design*, 65-80.

Anders, G. (2001). Pilot's attention allocation during approach and landing: Eye- and head-tracking research in an A330 full flight simulator. Paper presented at the *11th International Symposium on Aviation Psychology*, Columbus, OH.

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Banbury, S. & Tremblay, S. (Eds.) (2004). *A Cognitive Approach to Situation Awareness: Theory and Application*. Aldershot: Ashgate.

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.

Boag, C., Neal, A., Loft, S., & Halford, G.S. (2006). An analysis of the relational complexity in an air traffic control conflict detection task. *Ergonomics, 49*, 14, pp 1508-1526.

Boehm-Davis, D. A., Holt, R. W., Chong, R., & Hasberger, T. (2004). Using cognitive modeling to understand crew behavior. *Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA.

Boehm-Davis, D.A., Holt, R.W., Diez, M., & Hansberger, J.T. (2002). Developing and validating cockpit interventions based on cognitive modeling. In W.D. Gray, & C.D. Schunn, (Eds.) *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science*, p. 27.

Boucek Jr, G. P., Sandry-Garza, D. L., & Logan, A. L. Biferno, MA, Corwin, WH and Metalis, S.,(Douglas), 1987. In *Proceedings of the Workshop on the Assessment of Crew Workload Measurement Methods, Techniques, and Procedures: Part Task Simulation Data Summary, AFWAL-TR-87-3103, Sept* (pp. 15-16).

Buehler, R., Griffin, D., & Ross, M. (2002), Inside the planning fallacy: the causes and consequences of optimistic time predictions. In T. Gilovich, D. Griffin, & D Kahneman (Eds), *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, Cambridge, pp. 250-70.

Burdick, M.D., & Shively, R.J. (2000). A full-mission evaluation of a computational model of situational awareness. *Human Factors and Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: HFES

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*. CRC press.

Bzostek, J., Small, R., Bagnall, T., & Walters, B. (2005). *Intelligent Multimodal Signal Adaption System*. Micro Analysis & Design Final Report for NASA-Ames. Contract NNA05AC17C.

Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human–Computer Interaction*. Hillsdale, NJ: Erlbaum.

Carlin, A.S., Alexander, A.L, & Schurr, N. (2010). *Modeling pilot state in next generation aircraft alert systems*. Aptima, Inc.

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: CRC Press.

Corker, K. M., & Pisanich, G. (1998). Cognitive performance for multiple operators in complex dynamic airspace systems: Computational representation and empirical analyses. *Proceedings of the Human Factors and Ergonomics Society, 1*, 341-345.

Deutsch, S., & Pew, R. (2004). Examining new flight deck technology using human performance modeling. *Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA.

Deutsch, S.E., & Pew, R.W. (2008). D-OMAR: An architecture for modeling multitask behaviors. Chapter 8 in D.C. Foyle & B.L. Hooey (Eds.) *Human Performance Modeling in Aviation*. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 183-212.

Donath, D., & Schulte, A. (2009). Behavior model based recognition of critical pilot workload as trigger for cognitive operator assistance. In D. Harris (Ed.): *Engineering Psychology and Cognitive Ergonomics*, HCII 2009, LNAI 5639, pp. 518-528. Berlin / Heidelberg, Germany: Springer-Verlag.

Donnelly, D.M., Noyes, J.M., & Johnson, D.M. (1997). Decision making on the flight deck. *IEE Colloquium on Decision Making and Problem Solving*, pp.3/1-3/4, December 16.

Durso, F.T., Rawson, K., & Girotto, S. (2007). Comprehension and situation awareness. In F. T. Durso, R. Nickerson, S. Dumais, S. Lewandowsky, & T. Perfect, *Handbook of Applied Cognition (2nd),* Chicester: Wiley, pp. 163-193.

Elkind, J.I., Card, S.K., Hochberg, J., & Huey, B.M. (1990). *Human Performance Models for Computer-Aided Engineering*. New York, NY: Academic Press, Inc.

Endsley, M.R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society* 32nd Annual Meeting (Vol. 1, pp. 97-101). Santa Monica, CA: Human Factors Society.

Endsley, M.R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors, 37,* 65-84.

Eng, K., Lewis, R., Tollinger, I, Chu, A., Howes, A. & Vera, A. (2008) Generating automated predictions of behavior strategically adapted to specific performance objectives. *CHI 2008 Proceedings, Automatic Generation and Usability.* Montreal, Can.: Association for Computing Machinery.

Ericsson, K.A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102: 211-245.

Fleetwood, M.D., Lebiere, C., Archer, R., Mui, R., & Gosakan, M. (2006). Putting the brain in the box for human-system interface evaluation. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*. 1165-1169.

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: CRC Press.

Freed, M. (2000). Simulating human agents. *Papers from the 2000 AAAI Fall Symposium*, Michael Freed, Chair. Technical Report FS-00-03. Menlo Park, CA: AAAI Press.

Gil, G., Kaber, D., Kim, S., Kaufmann, K., Veil, T., & Picciano, P. (2009). Modeling pilot cognitive behavior for predicting performance and workload effects of cockpit automation. *Proceedings of the 2009 International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, 124-129.

Gil, G. H., Kaufmann, K., Kim, S. H., & Kaber, D. B. (2010). Effects of modes of cockpit automation on pilot performance and workload in a next generation flight concept of operation.

In *Proceedings of the 3rd International Conference on Applied Human Factors and Ergonomics* [CD-ROM]. Boca Raton, FL: CRC Press

Gil, G.H., & Kaber, D. (2012, in press). An accessible cognitive modeling tool for evaluation of pilot–automation interaction. *International Journal of Aviation Psychology, 22*.

Gonzales-Calleros, J., Vanderdonckt, J., Lüdtke, A. & Osterloh, J.P. (2010). Towards model-based AHMI development. *EICS '10*, June 21-23, Berlin, Germany.

Gore, B.F. (2008). Human Performance: Evaluating the Cognitive Aspects. *Handbook of Digital Human Modeling* (Ch. 32, pp. 1-18).

Gore, B.F. (2010). The use of behavior models for predicting complex operations. *Behavioral Representation in Modeling and Simulation* (BRIMS) 2010. Charleston, South Carolina, Simulation Interoperability Standards Organization (SISO): 1-4.

Gore, B.F. (2013, in prep). *The MIDAS User's Manual*. Moffett Field, CA: NASA Ames Research Center.

Gore, B.F. & Corker, K.M. (2000a). Human performance modeling: Identification of critical variables for national airspace safety. *Human Factors and Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: HFES.

Gore, B.F. & Corker, K.M., (2000b). Value of human performance cognitive predictions: A free flight integration application. *Human Factors and Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: HFES.

Gore, B. F., Hooey, B. L., Haan, N., Bakowski, D. L., & Mahlsted, E. (2011). A methodical approach for developing valid human performance models of flight deck operations. Paper presented at the *Human Computer Interaction International* (HCII) 2011, Orlando, FL.

Gore, B.F., Hooey, B.L., Mahlstedt, E., & Foyle, D.C. (2013). *Evaluating NextGen closely spaced parallel operations concepts with human performance models flight deck guidelines (Part 2 of 2)*. HCSL Technical Report (HCSL-13-02). Moffett Field, CA: NASA Ames Research Center.

Gore, B.F., Hooey, B.L., Mahlsted, E., & Foyle, D.C. (2012). Extending validated human performance models to explore NextGen Concepts. In S. Landry (ed.): *Advances in Human Aspects of Aviation*, pp. 407-416, Boca Raton, FL: CRC Press.

Gore, B. F., Hooey, B. L., Socash, C., Haan, N., Mahlsted, E., Bakowski, D. L., Gacy, A.M., Wickens, C.D., Gosakan, M., & Foyle, D. C. (2011). *Evaluating NextGen closely spaced parallel operations concepts with human performance models*. HCSL Technical Report (HCSL-11-01). Moffett Field, CA: NASA Ames Research Center.

Gore, B.F., Hooey, B.L., Wickens, C.D., & Scott-Nash, S. (2010). *A computational implementation of a human attention guiding mechanism in MIDAS v5.* In V.G. Duffy (Ed.): Digital Human Modeling, HCII 2009, LNCS 5620, pp. 237–246.

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*. Washington, D.C.: National Aeronautics and Space Administration. (NASA/CR-2010-216411).

Gore, B.F., Wickens, C.D., Hooey, B.L., Socash, C., & Gosakan, M. (2012), *MIDAS workload verification: Internal document*. Moffett Field, CA: NASA Ames Research Center.

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: A validation of GOMS for prediction and explanation of real-world task performance. *Human-Computer Interaction, 8*, 3, pp. 237-209.

Hart, S.G., Dahn, D., Atencio, A., & Dalal, K.M. (2001). Evaluation and application of MIDAS v2.0. In the *Proceedings of the Society of Automotive Engineers (SAE) World Aviation Congress*, Sept 2001, Seattle WA (SAE paper 2001-01-2648).

Hooey, B.L., & Foyle, D.C. (2008). Advancing the state of the art of human performance models to improve aviation safety. In D.C Foyle & B.L. Hooey (Eds.), *Human performance modeling in aviation* (pp. 321-349). Boca Raton, FL: CRC Press.

Hooey, B. L., Gore, B. F., Wickens, C. D., Salud, E., Scott-Nash, S., Socash, C., & Foyle, D. C. (2010). Modeling pilot situation awareness. Paper presented at the *Human Modeling of Assisted Technologies Workshop*, Belgirate, Italy.

Hooey, B.L., Gore, B.F., Mahlstedt, E., & Foyle, D.C. (2013*). Evaluating NextGen Closely Spaced Parallel Approach Concepts with Validated Human Performance Models Flight Deck Guideline*s (Part 1 of 2), HCSL Technical Report (HCSL-13-01). Moffett Field, CA: NASA Ames Research Center.

Horrey, W.J. & Wickens, C.D. (2004). Driving and side task performance: The effects of display clutter, separation, and modality. *Human Factors, 46*(4), 611-624.

Hüttig, G., Anders, G., & Tautz, A. (1999). Mode awareness in a modern glass cockpit– attention allocation to mode information. Paper presented at the *10th International Symposium on Aviation Psychology*, Columbus, OH.

John, B. E., Patton, E. W., Gray, W. D., & Morrison, D. F. (2012, September). Tools for Predicting the Duration and Variability of Skilled Performance without Skilled Performers. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 985-989). SAGE Publications.

John, B.E., Blackmon, M.H., Polson, P.G., Fennell, K. & Teo, L. (2009). Rapid theory prototyping: An example of an aviation task. *HFES 53rd Annual Meeting*. *53*(12), 794-798.

Joint Planning & Development Office. (June 2007). *Concept of operations for the next generation air transportation system*, *version 2.0,* Available online at: www.jpdo.gov/library/NextGen_v2.0.pdf, (accessed: 01/21/2009).

Karikawa, D., Takahashi, M., Ishibashi, A., Wakabayashi, T., & Kitamura, M. (2006). Human-machine system simulation for supporting the design and evaluation of reliable aircraft cockpit interface. *SICE-ICASE International Joint Conference*, pp.55-60.

Keller, J., Lebiere, C., & Shay, R. (2004). Cockpit system situational awareness modeling tool. In *Proceedings of the Human Performance, Situation Awareness and Automation Conference* (HPSAA II 2004), Daytona Beach, FL.

Klein, G., Orasanu, J., Calderwood, R., & Zsambok, C.E. (1993) *Decision Making in Action: Models and Methods*. Ablex Publishing Co., Norwood, NJ.

Kleinman, D.L., Baron, S. & Levison, W.H. (1971). A control theoretic approach to manned-vehicle systems analysis. *IEEE Trans on Auto Control*. Vol. AC-16, pp. 824-833, No. 6, December 1971.

Laird, J. E. (2008). Extending the soar cognitive architecture. In *Proceedings of the First Conference on Artificial General Intelligence (AGI-08).*

Laudemann, I., & Palmer, E. (1995). Quantitative analysis of observed workload in the measurement of aircrew performance. *International Journal of Aviation Psychology 5,* 187-197.

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*. Boca Raton, FL: CRC Press.

Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243-1270.

Lee, P.U., Sheridan, T., Poage, J.L., Martin, L., Jobe, K., & Cabrall, C. (2010). *Identification and Characterization of Key Human Performance Issues and Research in the Next Generation Air Transportation System (NextGen)*. NASA/CR-2010-216390.

Leiden, K., Laughery, K.R., Keller, J., French, J., Warwick, W., & Wood, S.D. (2001). *A Review of Human Performance Models for the Prediction of Human Error*. (Technical Report). Boulder, CO: Micro Analysis & Design, Inc.

Levison, W.H. (1989). The optimal control model for manually controlled systems. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R. Sutton, & L. Van Breda (Eds.) *Applications of Human Performance Models to System Design.* (Defense research series, Vol. 2). New York City, NY: Plenum Press. 185-200.

Levison, W., Elkind, J., & Ward, J. (1971). *Studies of multivariable manual control systems: A model for task interference*. NASA Contract report CR 1746. Washington, DC: NASA.

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*.

Lüdtke, A. & Osterloh, J-P. (2010). Modeling memory effects in the operation of advanced flight management systems. Paper presented at the *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.

Lyall, E.A. & Cooper, B., (1992). The impact of trends in complexity in the cockpit on flying skills and aircraft operation. *Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA.

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

McCarley, J., Wickens, C., Sebok, A., Steelman-Allen, K, Bzostek, J., & Koenecke, C. (2009). *Control of Attention: Modeling the Effects of Stimulus Characteristics, Task Demands, and Individual Differences*. NASA NRA: NNX07AV97A.

McNally, B.H. (2005). An approach to human behavior modeling in an air force simulation. *Proceedings of the 2005 Winter Simulation Conference*, pp.5, December.

McRuer, D.T., & Jex, H.R. (1967). A review of quasi-linear pilot models. *IEE Transactions of Human Factors in Electronics*, *HFE-8*(3), 231-249.

Miller, C.A. (1998). *Case Studies Involving W/Index*, Honeywell Technology Center.

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.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review,* 63 (2) 81–97.

Mumaw, R.J., Sarter, N.B., & Wickens, C.D. (2001). Analysis of pilots' monitoring and performance on an automated flight deck. *Proceedings of the 11th biennial meeting of the International Symposium on Aviation Psychology*, Dayton, OH: Wright State University.

Mumaw, R., Boorman, D.J., & Prada, R.L. (2006). Experimental evaluation of a new autoflight interface. Proceedings HCI-Aero 2006, *International Conference on Human Computer Interaction*, September 20-22, 2006, Seattle, Washington.

Muraoka, K., & Tsuda, H. (2006). Flight crew task reconstruction for flight data analysis program. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50*(11): 1194-1198.

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*.

North, R.A., & Riley, V.A. (1989). W/INDEX: A predictive model of operator workload. Applications of human performance models to system design. In G.R.B. McMillan, D.E. Salas. M.H. Strub, R. Sutton, L. van Breda (Eds.) *Applications of Human performance Models to System Design*. Defense Research Series. New York, Plenum Press. 2: 81-90.

Parks, D., & Boucek, G. (1989). Workload prediction, diagnosis and continuing challenges. In G.R. McMillan, D. Beevis ,E. Salas, M.H. Strub, R. Sutton, & L.van Breda. (1989). *Applications of Human performance Models to System Design*. Defense Research Series. New York, Plenum Press. 2.

Pew, R.W., & Mavor, A.S. (1998). *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington, DC: National Academy Press.

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

Polson, P. S., Irving, J., & Irving, S. (1994). *Applications of Formal methods of Human Computer Interaction to Training and Use of the Control And Display Unit*. Tech Report 94-08, University of Colorado.

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.

Raby, M., & Wickens, C.D. (1994). Strategic workload management and decision biases in aviation. *International Journal of Aviation Psychology*. Vol. 4, No. 3, pp. 211-240.

Reason J. (1990). *Human Error*. New York: Cambridge University Press.

Remington, R., Matessa, M., Freed, M., & Lee, S. (2003). Using Apex/CPM-GOMS to develop human-like software agents. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*. Melbourne: ACM Press.

Rickard, W.W., & Levison, W.H. (1981). Further tests of a model-based scheme for predicting pilot opinion ratings for large commercial transports. *Proceedings of the 17th Annual Conference on Manual Control*, pp. 247-256.

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., Young, M.S., Harris, D., Demagalski, J., Marshall, A., Waldmann, T., & Dekker, S. (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.

Salvucci, D.D., & Taatgen, N.A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, *115*, 101-130.

Sarno, K., & Wickens, C. (1995). The role of multiple resources in predicting time-sharing efficiency: An evaluation of three workload models in a multiple task setting. *International Journal of Aviation Psychology*, 5(1), 107-130.

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.

Schoelles. M.J., & Gray, W.D. (2011). Cognitive modeling as a tool for improving runway safety. *The Proceedings of the 16th International Symposium on Aviation Psychology*. Dayton, OH. 541-546.

Schoppek, W., & Boehm-Davis, D.A. (2004). Opportunities and challenges of modeling user behavior in complex real world tasks. MMI-Interaktiv, 7, June, 47-60. ISSN 1439-7854.

Schurr, N. (2011). ALARMS: Alerting and reasoning management system. *Presentation delivered to the 2011 NASA Aviation Safety Technical Meeting*, St. Louis, MO.

Sebok, A., Wickens, C., Leiden, K., Kamienski, J., & Bagnall, T. (2006). *Cockpit-Based Wake Vortex Visualization: Final Report*. Contract No. NNL06AA28P, NASA Langley.

Sebok, A., Wickens, C., Sarter, N., Quesada, S., Socash, C., & Anthony, B. (2012). The automation design advisor tool (ADAT): Development and validation of a model-based tool to support flight deck automation design for nextgen operations. *Human Factors and Ergonomics in Manufacturing and Service Industries, 22*(5), 378-394.

See, J.E., & Vidulich, M.A. (1998). Computer modeling of operator mental workload and situational awareness in simulated air-to-ground combat: An assessment of predictive validity. *The International Journal of Aviation Psychology, 8*(4), 351-375.

Sherry, L., Polson, P., Feary, M., & Palmer, E. (2002) *When Does the MCDU Interface Work Well? Lessons Learned for the Design of New Flightdeck User-Interfaces*. Honeywell Publication C69-5370-0021.

Shively, R. J., Brickner, M., & Silbiger, J. (1997). A computational model of situational awareness instantiated in MIDAS. *Proceedings of the Ninth International Symposium on Aviation Psychology*, Columbus, Ohio.

Stanton, N.A., Salmon, P., Harris, D., Demagalski, J., Marshall, A., Waldmann, T. & Dekker, S. (2003). Predicting pilot error: Assessing the performance of SHERPA. *Proceedings of the HCII Conference*, 587-591.

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

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.

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*.

Stone, G., Culick, R., & Gabriel, R. (1987) Use of task timeline analysis to assess crew workload. In A. Roscoe (Ed) *The practical assessment of pilot workload*. NATO AGARDograph #282.

Sulistyawati, K., Wickens, C.D., & Chui, Y.P. (2011). Prediction in Situation Awareness: confidence bias and underlying cognitive abilities. *The International Journal of Aviation Psychology*, *2*(2), 153-174.

Svensson, E.A.I., & Wilson, G.F. (2002). Psychological and Psychophysiological Models of Pilot Performance for Systems Development and Mission Evaluation. *The International Journal of Aviation Psychology, 12*(1), 95-110.

Tidhar, G., C. Heinze, & Selvestrel, M. (1998). Flying together: modeling air mission teams. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 8*(3), 195-218.

Uijtde Haag, M., Duan, P., Schnell, T., Cover, M., Anderson, N., Snow, M., Etherington, T., Rademaker, R., & Theunissen, E. (2011). *Hazard and integrity monitoring and integrated alerting and notification methods*. Presentation delivered to the 2011 NASA Aviation Safety Technical Meeting in St. Louis, MO.

Walden, R.S., & Rouse, W.B. (1978). A Queueing Model of Pilot Decisionmaking in a Multitask Flight Management Situation. *IEEE Transactions on Systems, Man and Cybernetics* (pp.867-875), December 1978.

Wickens, C.D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and performance* (Vol. 7, pp. 239–257). Hillsdale, NJ: Erlbaum.

Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of Attention* (pp. 63-101). New York: Academic Press.

Wickens, C.D. (1986). The effects of control dynamics on performance. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of Perception and Performance Vol. II* (pp. 39-1/39-60). New York: Wiley & Sons.

Wickens, C.D. (1990). Resource management and time-sharing. In J.I. Elkind, S.K. Card, J. Hochberg, & B.M. Huey (Eds.), *Human performance models for computer-aided engineering* (pp. 181-202). Orlando, FL: Academic Press.

Wickens, C.D. (2002a). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177.

Wickens, C.D. (2002b). Situation awareness and workload in aviation. *Current Directions in Psychological Science*, *11*(4), 128-133

Wickens, C.D. (2005). Multiple resource time sharing models. In N. Stanton et al. (Eds.), *Handbook of human factors and ergonomics methods* (pp. 40-1/40-7). Boca Raton, FL: CRC Press.

Wickens, C.D. (2008a). Situation awareness. Review of Mica Endsley's articles on situation awareness. *Human Factors, Golden Anniversary Special Issue, 50*, 397-403.

Wickens, C.D. (2008b). Multiple resources and mental workload. *Human Factors Golden Anniversary Special Issue*, *3*, 449–455.

Wickens, C.D., Bagnall, T., Gosakan, M., & Walters, B. (2011). A cognitive model of the control of unmanned aerial vehicles. *The Proceedings of the 16th International Symposium on Aviation Psychology*, Dayton, OH, 535-540.

Wickens, C.D., Gempler, K., & Morphew, M.E. (2000). Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors Journal,* 2(2), 99-126.

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

Wickens, C.D., Harwood, K., Segal, L., Tkalcevic, I., & Sherman, B. (1988). TASKILLAN: A simulation to predict the validity of multiple resource models of aviation workload. *Proceedings of the 32nd Meeting of the Human Factors Society.* Santa Monica, CA: Human Factors Society, 168-172.

Wickens, C.D., Hooey, B.L., Gore, B.F., Sebok, A., & Koenecke, C.S. (2009). *Identifying black swans in nextgen: Predicting human performance in off-nominal conditions*. *Human Factors. 51*(5), 638-651.

Wickens, C.D., Larish, I., & Contoror, A. (1989). Predictive performance models and multiple task performance. *Proceedings of the Human Factors Society 33rd Annual Meeting*, pp 96-100.

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 (A/SA) model of pilot error. In D.C. Foyle & B.L. Hooey (Eds.) *Human Performance Modeling in Aviation*. CRC press.

Wickens, C.D., Sandry, D., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, output, and central processing. *Human Factors*, *25*, 227-248.
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.
Wickens, C.D., Vincow, M.A., Schopper, A.W., & Lincoln, J.E. (1997). Computational models of human performance in the design and layout of controls. *Crew System Ergonomics Information Analysis Center, CSERIAC 97-02.* Dayton, OH: Wright-Patterson Air Force Base.
Yeh, Y., & Wickens, C.D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30* (1): 111-120.
Zacharias, G. L., Miao, A. X., Illgen, C., Yara, J.M, & Siouris, G.M. (1996). SAMPLE: Situation awareness model for pilot in the loop evaluation. *Conference on Situation Awareness in the Tactical Environment*, Naval Air Warfare Center, Patuxent River, MD.

8. APPENDICES

A. MODEL VALIDATION DETAILS (EXTRACT FROM CSERIAC REPORT)
B. REFERENCES INCLUDED IN THE MODEL EVALUATIONS

Appendix A: CSERIAC Report Excerpt

This material is a modified version of Chapter 1 of Wickens et al, 1997.

1. Why Model?

1.1 Early Design Decisions

It is well established that implementing design changes in a system once it has gone into production is very difficult, because of the high cost involved. Thus, human factors deficiencies in a system that are revealed through product evaluation are often discovered too late to be corrected (Elkind, Card, Hochberg, & Huey, 1990). Human factors input to the design process must be provided early in the design cycle, before extensive production setup costs are incurred. But, in the absence of an existing system or prototype configured with a human in the loop to evaluate the design, where should such input originate? A strong case can be made for using computer-based models to "compute" the efficiency of proposed designs from a human factors standpoint (Corker & Smith, 1993). While such models may not be likely to indicate the "best" design, they should be able to raise a red flag by revealing important human factors deficiencies when they are run using a representation of the proposed design as model input. The sorts of human factors deficiencies that directly concern us here are, of course, deficiencies in the layout of information appearing on a display, and, when appropriate, the physical placement of displays and associated controls.

1.2 Conflicting Principles

As those working in the design community will testify, it is rare that a given display arrangement satisfies all human factors layout principles simultaneously. For example, moving one frequently used display close to another frequently used display may mean placing it *farther* away from a display to which it is related. Placing related displays close together may create problems if they are placed *so* close together that the layout becomes cluttered. When two (or more) principles conflict, which one should be followed? Is the best solution a compromise that violates both principles to some extent? A validated model can incorporate answers to such questions and guide the display layout process. A model that takes account of the various principles and provides weights characterizing their relative importance can provide a number that expresses the added benefit of adhering to (or cost of violating) a combination of principles for a given proposed design. Then, the optimal design (or best compromise) can be predicted.

For example, if adhering to principle A (e.g., moving a display closer to another display used in sequence) is twice as important (leads to twice the predicted performance gain) as adhering to principle B (e.g., moving the display closer to a functionally related control), then there is solid justification for choosing a design solution that conforms to A but violates B, even if cost or engineering factors may slightly favor B. In the absence of such models, the designer is left in a state of frustration when a list of sometimes conflicting principles is offered, but no guidance is given on how to resolve these conflicts except by doing further costly experiments. In such instances, it is understandable that the designer will simply choose to satisfy the principles that can be applied most economically. The availability of models should help to address this situation.

1.3 Figure of Merit

A third reason why computational models are important is that they make it possible to assess the degree to which a display layout adheres to human factors principles by providing the designer with

a predicted *figure of merit* for a given display layout. When there is a computational algorithm for measuring the strength of adherence to (or violation of) each principle in isolation, then the overall "goodness" of a display layout can be assessed by properly combining these assessments of adherence to each principle. This combination should, of course, weight the degree of adherence to a given principle by the degree of *importance* of that principle to system performance. In this way, different display layouts can be compared and evaluated before manufacturing begins to determine which one is the "best." Or, alternatively, the impact of a particular design decision (e.g., to reposition a display to a side panel) can be evaluated and its cost or benefit expressed in quantitative terms. Such quantitative comparative data should be useful in trading off human factors constraints against other engineering design or cost constraints. As we have noted, it is particularly valuable to have these data in advance of the actual manufacturing process (i.e., before "metal is bent"), because of the incredibly high cost of making adjustments in design later to accommodate human factors concerns that were revealed only after production had begun (Elkind et al., 1990).

2. Computational Models and Measures: Properties and Criteria

By a computational model, we mean here a tool that "computes," from inputs of parameters that reflect characteristics of a pilot-task-interface description, some numerical index of the quality of performance of the pilot-aircraft system. Such computation is typically (but need not be) done on a computer, in that certain models can predict outputs via a simple algebraic formula (Elkind et al., 1990). A high figure of merit based on a validated model predicts relatively good performance in terms of accuracy and/or speed &/or workload &/or situation awareness. At a minimum, the model should make ordinal predictions (A is better than B is better than C). Preferably, it should be able to make interval- or ratio-scale predictions of how much better A is than B.

Many models are based upon measures of two sorts. There are measures that may define the input to a model, such as the complexity of a particular flight deck operation, or the computed salience of an alert. There are also measures of the model output, performance, workload and/or situation awareness, and these measures, to be useful in model validation, must often be operationalized in particular procedures (e.g, NASA TLX measures of workload; relational complexity as a measure of cognitive complexity. A measure does not become a model, however, until it is incorporated into a quantitative expression that predicts the direction and approximate magnitude of the effects on performance.

There are a number of important criteria by which the value of a model or measure can be judged. The following sections describe the four most important of these criteria: **validation, complexity, practical significance, and a priori specification**. A fifth criterion, usability, is also mentioned, but it is more difficult to apply to the models reviewed in this report.

2.1 Validation

Validating a computational model or measure is very similar in many respects to validating a test. The goal of test validation is to determine if the score earned by an individual on the test correlates with or predicts some *criterion* score measured under other, usually more "operational," circumstances (Anastasi, 1988; Allen & Yen, 1979). For a pilot performance model we ask whether the score (level of performance) predicted for a particular pilot-task-interface combination correlates with the level of performance measured under operational conditions. Do pilots actually perform better in the conditions that are predicted to be better? Do pilots make the same kinds of errors that are predicted?

While we will see that few pilot performance models have been validated in real-world conditions, many have not even been validated against human performance at all. This fact is not meant to be a criticism of their developers; rather, it should serve as a caution to potential users to seek validation of the and also to evaluate the *quality* of whatever validation has been done. In the paragraphs below, we outline some of the factors that should be considered in assessing the quality of empirical validation.

To provide some concrete context for this discussion, suppose a model is developed to predict the optimal layout of cockpit instruments for aircraft flight (Andre & Wickens, 1991). Eight possible layouts are constructed, and the model incorporates parameters that characterize each layout in order to predict a level of performance for each. Pilots then perform a flight task with each layout, a performance score is derived, and a *correlation* is used to evaluate the extent to which the model predictions for each layout match with the obtained data (Figure 1a). The higher the correlation, the better the model and the more successful is the validation. That is, if a model has had its predictions matched against performance, we can say it has been validated. To the extent that the correlation returned by such matching is high, we can say that the validation is **successful.** There are four critical elements involved in creating this correlation and using it to demonstrate model validation: the criterion, the operator sample, the condition sample, and the statistics. We discuss each of these below, to show how their characteristics can influence the validity of the model.

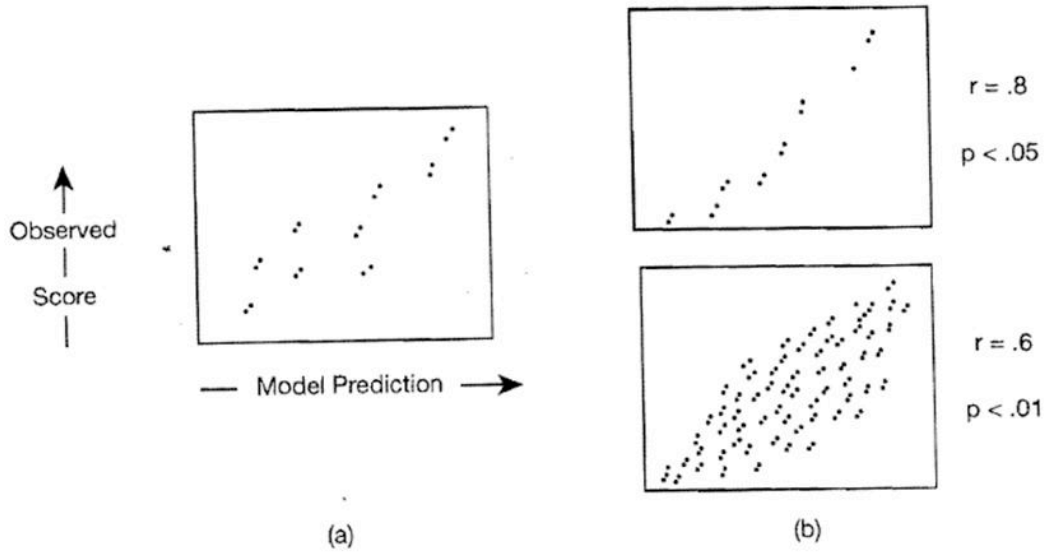

Figure 1. Correlation in model validation.

2.1.1 The Criterion. The criterion variable is the variable that is measured on the y-axis of Figure la. In the Andre & Wickens example, it is a measure of how good a job the display layout does in supporting performance of the task. But what does "good" mean? Ideally, the variable used should be some measure of performance evaluated against objective criteria. In our example, this might include reduced flight path deviation or more rapid and accurate hazard detection. These criterion variables are preferred over operator opinion, simply because in so many circumstances, what users say about an interface does not necessarily agree with how they perform using the interface (Andre & Wickens, 1995).

Ideally, the criterion variable should also be measured in more operational circumstances. We argue that, all other things being equal, measurement in the aircraft is more valid than measurement in the simulator, and measurement in the simulator is more valid than measurement in the basic laboratory setting. But these suggestions do not mean that laboratory studies, or those that have used operator opinion to solicit data on the ordinate of Figure 1a, are "invalid." Our argument is simply that they will be *less robust* than simulator-based and/or real-world measures, all other factors being equal. These "other factors," which are sometimes traded off in favor of more basic laboratory validation, are related to the sample, as discussed next.

The Sample. Each correlation is based upon a set of data points or *cases* (e.g., the points in Figure 1a, which is called the *sample).* The measure of validation is based in part on the identity of those cases. Generally, the greater the degree to which the people whose performance is evaluated are *typical of the real world users* of the system, the greater the generalizability to the actual applied setting. Hence, testing pilots would provide a more meaningful evaluation of a model of cockpit display layout than testing non-pilots.

One reason that validation studies sometimes fail to employ typical system users (experts) is the lack of availability of those experts, which may make it difficult to obtain a *sufficient sample size.* When sample size is small, a correlation can appear to be very high, yet not be statistically reliable, in the sense that it may not be replicated in future studies. Validation studies should seek a large sample, and reports of these studies should always indicate the sample size. Note that if the users are in scarce supply and the real world conditions are expensive to create, it is the need for large samples that sometimes leads model validation efforts away from using typical users and real world conditions.

As noted, the sample is made up of cases. The reported correlation will differ depending on whether those cases, the individual data points in Figure 1a, each represent data that are (a) the average over subjects of performance in each of several test conditions, (b) the individual measure of each subject in each condition, or (c) the average for each subject over several test conditions. Since models are typically developed to predict the effect of different conditions, rather than differences among users, the third option is not appropriate, and is rarely used. The choice between the first two options however is not trivial, however, and which method is used should be clearly reported when correlations are presented.

To illustrate the consequences of this choice, Figure 2 shows validation of a model on four test conditions (A-D) with three subjects (1-3). The raw data points used to compute correlations can represent either the means for the different conditions (averaged over subjects, Figure 2a), the means for the different subjects (averaged over conditions, Figure 2c), or a heterogeneous mixture of the means for each subject and condition (center panel, Figure 2b).

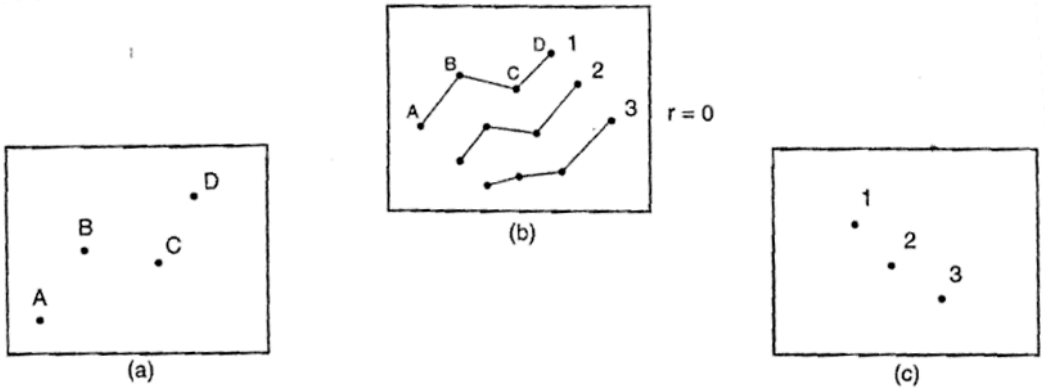

Figure 2. Results of validation study testing three subjects (1-3) under four conditions (A-D). (a) Illustrates a positive correlation between model prediction and obtained data for the four conditions (averaged over subjects). (b) Shows a 0 correlation when variance between subjects is combined with variance between conditions. (c) Shows a negative correlation when data for each subject are averaged over conditions.

As the figure illustrates, each technique can yield a quite different correlation between predicted and obtained scores. Nevertheless, model validation studies are not always explicit about which form of correlation is used. We suggest that the first technique is of greatest value in model validation because it removes variance accounted for by individual differences among subjects, which the model is generally not intended to predict. A plausible alternative is to compute the correlation between model prediction and performance for each subject individually (e.g., correlation underlying the 4 points on each line in Figure 2b) and then report the average correlation across subjects (See Wickens, McCarley, Alexander, Thomas, Ambinder, & Zheng, 2008 for an example). In many respects this approach is optimal because it both shows how well the model fits individual pilot data, as well as the data of the "mean pilot".

2.1.2 The Conditions. Model validation requires the construction of a set of conditions, like the eight display layouts described above, that vary along some parameter(s) incorporated in the model. It is important that the *range of the parameters be appropriate*., neither two wide nor too narrow. Validation efforts may sometimes fall short by creating too little variance between the conditions, relative to the power of the model. For example, if one parameter in the model relates to mean display *separation*, and the validation study used four separations varying in increments of only 0.5 cm, the model might predict little variance in performance. With little variance predicted, a correlation with performance cannot be expected to rise very high, and the validation effort is doomed to failure.

At the other extreme, it is easy to select two or three cases to evaluate that differ by such obvious and excessive magnitudes that variance in performance between them is virtually guaranteed. To take our previous example, we might select display separation differences of 1, 10, and 100 cm. In this case, the differences in visual scanning requirements are large enough to guarantee substantial performance differences (probably lower performance with the wide separation). In this regard, it should also be noted that a single "outlier" point in the appropriate corner of a scatter plot can greatly inflate (Figure 3a) or deflate (Figure 3b) the value of a correlation, particularly if the sample is small, and hence make it appear that the model is doing a much better (or worse) job of predicting performance than it really is. This is illustrated in Figure 3. In 3a, the model appears to have no predictability across the four conditions at the lower left, yet because of the single outlier, the

correlation will be high. In Figure 3, the model does a very good job predicting variance in performance across the four conditions in the lower right, but because of the outlier, the correlation would be very low, and perhaps negative.

The above concerns can be addressed by carefully selecting the parameter values to span the range to be considered by the model (or to be realistically incorporated in actual system design), and, ideally, **by presenting the raw scatter plot** from which the correlations are derived and carefully labeling or identifying any "outlier" conditions that might predict one extreme reading. Correlations should be reported both with and without the outlier, and a clear discussion is needed as to the possible reasons for its removal. That is, what **unique** properties of the condition in question might have caused its displacement from the rest of the data points.

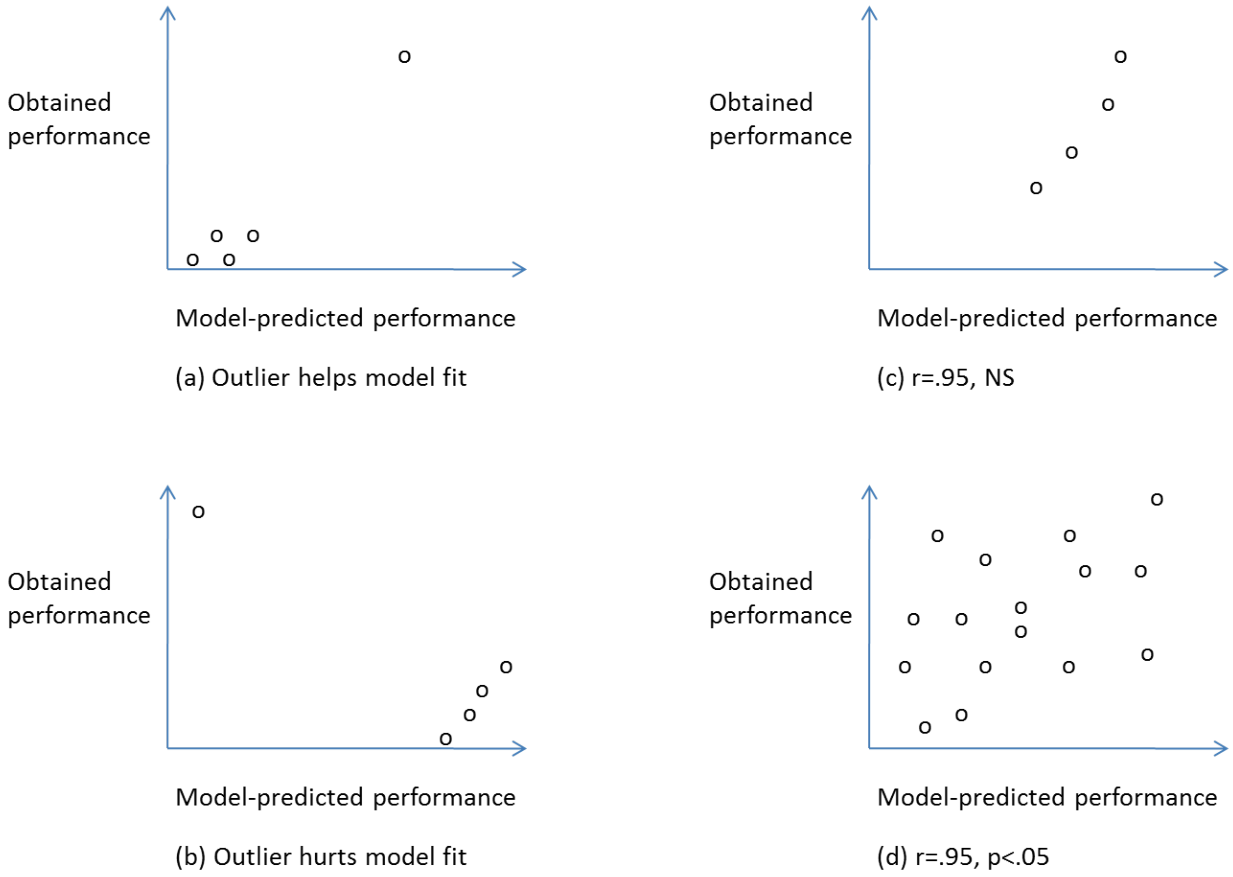

Figure 3. Four examples of model prediction (X axis) vs. obtained performance (Y axis). (See text for a description. The correlation values shown in the figures are not precisely accurate, but are illustrative)

2.1.3 The Statistics. Within the context of a model, the correlation coefficient is a statistic that is often used to characterize the relationship between predicted and obtained scores. The Spearman product moment correlation is the statistic typically used, but rank-order correlations are also sometimes appropriate (particularly when extreme values, as discussed above, may inflate the Spearman value).

Three aspects of the correlation are important to validate a model across conditions: its value (from -1.0 to 0 to +1.0) and its significance (based jointly on its value and the sample size, N, or the number of points in the scatter plot), and N. In classical statistics, most weight is given to significance. But in model validation, equal concern should be given to the value of r. This is illustrated in figure 3c and d. In figure3c the model does a beautiful job of predicting performance, but the high correlation (0.95) may not reach "significance" simply because of the small N. In figure 3d the model is not very good at predicting performance, the correlation is low but it may be significant because of the much larger value of N. Hence authors should always report both the correlation value and its significance and N.

A special case in the statistics of model validation, which is important because of its prominence in many of the models discussed in this report, concerns *linear regression models,* in which the parameters for the model are themselves derived from empirical data. In a typical linear regression approach, the eight display layouts in our experiment might each be characterized by their quantitative level along each of three display variables captured by a model parameter (e.g., average separation, degree of clustering by relatedness, degree of importance). After the dependent variable (e.g., a performance measure) is assessed, linear regression will provide weights dictating the relative importance of each of the three parameters in accounting for variance across the criterion measure (M), such that:

$$M=aA+bB+cC$$

where *A, B,* and C are the levels of each of the three parameters; and *a, b,* and c are the weights of each parameter in the equation.

A well-known characteristic of multiple regression models is that they tend to provide an overly optimistic picture of how well the parameters do in predicting the model dependent variable, because these predictions will capitalize on, and be able to account for, chance or random variation in the criterion measure. This problem can be dealt with in three ways. First, there are objective techniques of "correction for shrinkage" (Tatsuoka, 1971) that reduce the stated estimate of the predicted variance accounted for. Second, and more preferable, is *cross-validation.* In one form of cross-validation, the regression weights are derived on one sample (of conditions, subjects, or both), and then applied to a different sample to determine how well the data of the latter sample are predicted. The latter is then used as the reported validation measure. In another form of cross-validation, the validation experiment is simply repeated. Finally, the third defense against the problem, which is compatible with either of the first two, is to restrict the parameters in a marketed model to those that "make sense" in terms of a conceptual model of human information processing. That is, a parameter might describe a component like working memory load or the breadth of attention that has an independently validated role in human psychology.

All three of these corrective procedures tend to reduce the amount of "significant variance" accounted for by a model and hence, in a way, make the model appear less effective. For this reason, and because of the added complexity of carrying out the first two defenses, many developers of multiple regression models may be reluctant to apply the procedures. But potential model users should be aware of these possible constraints on the validity of regression-based models.

2.1.4 Qualitative Validation. Finally, we note a form of validation in qualitative rather than quantitative form. Here the model predicts some pattern of effects, or distribution of error types, or a

particular error occurring in a particular circumstance. There is no number describing the level of successful prediction, but only data presented that suggest a matching pattern.

3. Complexity

Complexity is a second feature that characterizes the potential value of a model to the user. This feature distinguishes single parameter from multi-parameter models.

Single parameter models may do a precise job of predicting the effect of a given variable (e.g., spatial separation, location in the visual field) on display processing, but they are less than fully satisfactory for use in design. The reason is that they fail to address how the impact of the given parameter will be influenced by other environmental or task characteristics that may vary in real-world conditions. The influence of one model parameter on performance may be greatly affected by (i.e., interact with) another parameter. For example, changes in the spatial separation between two displays will have a very different impact when the displays are related to the same task than when they are not. Hence, it is desirable that models not only recognize the need to address multiple (pertinent) predictive variables but also to include higher-order terms to examine and account for the impact of their potential interactions.

4. Practical Significance of Effects

The value of a model or measure to the user also depends on whether the predictions of the model (or the effects due to the characteristic measured) are on a scale that makes a difference in practice. In operational circumstances, time differences of seconds, or sometimes as short as tenths of a second, are generally of practical importance. In contrast, differences on the order of milliseconds rarely have operational relevance in any sense other than a theoretical one, even though under careful laboratory control small differences may take on a high degree of statistical significance. The magnitude of time difference that is important is, to some extent, context dependent, however. For models predicting the time to perform highly predictable and repetitive operations like key strokes (Card, Moran, & Newell, 1986) differences of tenths or even hundredths of a second can be meaningful.

In general, then, the greater the magnitude of the typical time effects predicted by the model, the greater the value of the model in terms of its *practical significance.* Models that predict errors also score highly on the practical significance criterion.

Note that many validated effects in the millisecond range as demonstrated in the controlled laboratory may indeed turn out to have important design implications. But the *robustness* of these effects in more complex, less controlled environments should be assessed before certifying them as important components of display layout models. Many theoretically important models of visual attention processes have been intentionally excluded from consideration in this report because their practical significance has been deemed low.

5. A Priori Specification

This criterion assesses the degree to which the user of the model or measure can specify model parameters or compute the measure without collecting any new data. For example, some models

require that frequency-of-use estimates for different display components or transition probabilities between displays be obtained before computation of layout design measures can begin (e.g., Freund & Sadosky, 1967; Wierwille, 1981). This feature makes the models more difficult to use. Models and measures that require only information about the physical characteristics of a display are easier to apply than models and measures that require specification of aspects of the task or display content. Models that require formal experiments or observational studies to obtain values for model parameters are of limited usefulness to many designers.

6. Usability

One final criterion that should not be overlooked is the usability of the model or measure for the designer to whom it is recommended. Rouse and Cody (1989) note that designers' use of data sources is heavily dependent on the ease of using those sources. Hence, a well validated model that incorporates many variables, predicts significant performance differences, and requires no data collection to implement may nevertheless be neglected because its features simply make it too difficult for the non-expert to use. Ironically, sometimes the model builder's quest for high validity and explanatory power can make the model so complex (i.e., with too many parameters that must be specified) that it will remain unused. Other features affect usability as well, however. Good human factors at the computer interface, well-written instruction manuals, understanding of the model user's domain, and compatibility with readily available hardware are all examples. While we consider the usability criterion to be critical, it is difficult to apply to most of the models reviewed here, since they are not sufficiently mature for formal computer-based software and user interfaces to have been developed.

7. References

Allen, M.J., & Yen, W.M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks / Cole Publishing Company.

Anastasi, A. (1988). *Psychological Testing* (6th Edition). New York, NY: Macmillan.

Andre, A.D., & Wickens, C.D. (1991). *A computational approach to display layout analysis*. (Technical Report ARL-91-6/NASA-91-2). Savoy, IL: University of Illinois, Institute of Aviation, Aviation Research Laboratory.

Andre, A.D., & Wickens, C.D. (1995). When users want what's not best for them: A review of performance-preference dissociations. *Ergonomics in Design*. October, 10-14.

Card, S.K., Moran, T.P., & Newell, A. (1986). The model human processor: An engineering model of human performance. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.) *Handbook of Perception and Human Performance* (Vol. II, pp 45-1 to 45-35). New York, NY: Wiley.

Corker, K.M., & Smith, B.R. (1993). An Architecture and Model for Cognitive Engineering Simulation Analysis: Application to Advanced Aviation Automation. Paper presented at the *AIAA Computing in Aerospace 9 Conference*, San Diego, CA.

Elkind, J.I., Card, S.K., Hochberg, J., & Huey, B.M. (1990). *Human Performance Models for Computer-Aided Engineering*. New York, NY: Academic Press, Inc.

Freund, L.E., & Sadosky, T.L. (1967). Linear programming applied to optimization of instrument panel and workplace layout. *Human Factors*, *9*, 295-300.

Rouse, W.B., & Cody, W.J. (1989). Designers' criteria for choosing human performance models. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R.Sutton, & L. Van Breda (Eds.), *Applications of Human Performance Models to System Design*. (pp 7-14). New York, NY: Plenum Press.

Tatsuoka, M.M. (1971). *Multivariate Analysis: Techniques for Educational and Psychological Research*. New York, NY: Wiley.

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. Chapter 9 in D.C. Foyle and B.L. Hooey (Eds.) *Human Performance Modeling in Aviation*. Boca Raton, FL: CRC Press.

Wierwille, W.W. (1981). Statistical techniques for instrument panel arrangement. In J. Moraal & K. Kraiss (Eds.), *Manned Systems Design* (pp. 201-281). New York, NY: Plenum.

Appendix B: Sources Included in the Model Evaluations

Al-Zubaidy, S.N. (2008). Proposal for modeling the piloting system. *Second Asia International Conference on Modeling & Simulation*, AICMS, pp.800-805, May 13-15.

Anderson, M.R., Clark, C., & Dungan, G. (1995). Flight test maneuver design using a skill- and rule-based pilot model. *IEEE International Conference on Systems, Man and Cybernetics, Intelligent Systems for the 21st Century*, October 22-25, pp. 2682-2687.

Anderson, M., & Schmidt, D. (1985). Closed-Loop Pilot/Vehicle Analysis of the Approach and Landing Task, pp. 522-526.

Andrews, J.A. (1991). Unalerted air-to-air visual acquisition. Technical report under Air Force contract F19628-90-C-0002. DOT/FAA/PM-87/34, NTIS N9213577. Lexington, MA: MIT Lincoln Laboratory.

Banbury, S., & Tremblay, S. (2004). *A cognitive approach to situation awareness: theory and application*. Ashgate Pub Limited.

Barcheus, F., Ulfvengren, P., & Martensson, L. (2010). Communication enablers for delegation - A relational model for the new ATM system. Paper presented at the *Human Computer Interaction Aero Conference 2010*, Cape Canaveral, FL.

Barnett, B., Stokes, A., Wickens, C.D., Davis, T. Rosenblum, R., & Hyman, F. (1987). A componential analysis of pilot decision-making. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA: HFES.

Baron, S. (1983). An optimal control model analysis op data from a simulated hover task. Proceedings of the *18th Annual Conference on Manual Control*, pp. 195-215.

Baron, S. & Corker, K. (1989) Engineering-based approaches to human performance modeling. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R. Sutton, and L. van Breda (Eds.) *Applications of Human Performance Models to System Design*. (Defense research series, Vol. 2). New York City, NY: Plenum Press. Pp. 203–217

Baron, S., Zacharias, G., Muralidharan, R., Huraldiharan, & Lancraft, R. (1980). Procru: a model for analyzing flight crew procedures in approach to landing. *Proceedings of the 16th Annual Conference on Manual Control*, pp. 495-526.

Bautsch, H., McNeese M. D., & Narayanan, S. (1997). Assessing the value of human performance modeling in exploring pilot-system dynamics. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting.* Santa Monica, CA.

Belyavin, A.J. & Spencer, M.B., (2004). Modeling performance and alertness: The QinetiQ approach. *Aviat Space Environ Med; 75*(3, Suppl.): A93–103.

Belyavin, A., Woodward, A., Nguyen, D., Robel, G., & Woolworth, J. (2005). Development of a novel model of pilot control behavior in balked landings. In *2005 AIAA Modeling and Simulation Technologies Conference and Exhibit* (pp. 1-11).

Benjamin, P. (1970). A Hierarchical Model of a Helicopter Pilot. Human Factors, 12, 361-374.

Besco, R. (1988). Modelling system design components of pilot error. Society of Automotive Engineers Technical Papers. Warrendale, PA. Paper 872517, pp. 53 - 58.

Best, B., Lebiere, C., Schunk, D., Johnson, I., Archer, R. (2005). Validating a Cognitive Model of Approach based on the ACT-R Architecture

Blom, H., Corker, K., Stroeve, S., & Van Der Park, M. (2003). Study on the integration of Air-MIDAS and TOPAZ (NLR-CR-2003). San Jose, CA: NASA/ATAC Corp.

Blom, H.A.P., Corker, K.M., & Stroeve, S.H. (2005). On the Integration of Human Performance and Collision Risk Simulation Models of Runway Operation. In the *6th USA/Europe Air Traffic Management R&D Seminar*, Baltimore, USA, 27-30th June 2005. pp 1-10.

Boehm-Davis, D. A., Holt, R. W., Chong, R., & Hasberger, T. (2004). Using cognitive modeling to understand crew behavior. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting.* Santa Monica, CA.

Boehm-Davis, D.A., Holt, R.W., Diez, M., & Hansberger, J.T. (2002). Developing and validating cockpit interventions based on cognitive modeling. In W.D. Gray & C.D. Schunn (Eds.) *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science*, p. 27

Broussard, J.R., & Stengel, R.F. (1977). Modern control analysis of the pilot-aircraft system. *IEEE Conference on Decision and Control including the 16th Symposium on Adaptive Processes and A Special Symposium on Fuzzy Set Theory and Applications*, pp.235-240, December 1977.

Burdick, M.D., & Shively, R.J. (2000). A full-mission evaluation of a computational model Of situational awareness. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA.

Byrne, M.D., & Kirlik, A.(2004). Integrated Modeling of Cognition and the information environment: A Closed-Loop, ACT-R Approach to Modeling Approach and Lanading With and Without Synthetic Vision System (SVS) Technology.

Byrne, M.D., Kirlik, A., & Fleetwood, M.D. (2008). An ACT-R approach to closing the loop on computational cognitive modeling. Chapter 5 in D.C. Foyle & B.L. Hooey (Eds.) *Human Performance Modeling in Aviation*. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 77 - 104.

Callantine, T.J., (2003). Detecting and Simulating Pilot Errors for Safety Enhancement. SAE Technical Papers.

Carbonell, J. R. (1966). A queueing model of many-instrument visual sampling. *Human Factors in Electronics, IEEE Transactions on*, (4), 157-164

Carlin, A.S., Alexander, A.L., & Schurr, N. (2010). Modeling pilot state in next generation aircraft alert systems. Aptima, Inc.

Chunguang, W., Feng, L., Junwei, H., & Guixian, L. (2008). A Revised Optimal Control Pilot Model for Computer Simulation. The *IEEE International Conference on Bioinformatics and Biomedical Engineering (ICBBE)*, May 16-18

Colle, H.A. & Reid, G.B. (2005). Estimating a mental workload redline in a simulated air-to-ground combat mission. *The International Journal of Aviation Psychology*, *15*(4), 303-319.

Colvin, K., Funk, K., & Braune, R. (2005). Task Prioritization Factors: Two Part-Task Simulator Studies. *The International Journal of Aviation Psychology*, *15*(4), 321-338.

Camacho, R. (1995). Using Machine Learning to extract models of human control skill. *Proceedings of AIT'95*.

Corker, K.M. (2000). Cognitive models and control: human and system dynamics in advanced airspace operations. In N.B. Sarter & R. Amalberti (Eds.) *Cognitive Engineering in the Aviation Domain*. Mahwah, NJ: Lawrence Erlbaum Associates. Pp. 13-42.

Corker, K., H.A.P. Blom, S.H., & Stroeve (2005). Study on the integration of human performance and accident risk assessment models: Air-MIDAS & TOPAZ. *Proceedings of the 2005 International Symposium on Aviation Psychology.* Dayton, OH: Wright State University, 147-152.

Corker, K.M., Muraoka, K., Verma, S., Jadhav, A., & Gore, B.F. (2008). Air MIDAS: A Closed-Loop Model Framework. Chapter 7 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 145-182.

Corker, K.M., & Pisanich, G.M. (1995). Analysis and modeling of flight crew performance in automated air traffic management systems, Oxford, UK, Pergamon.

Corker, K. M., & Pisanich, G.M. (1998). Cognitive performance for multiple operators in complex dynamic airspace systems: Computational representation and empirical analyses. *Proceedings of the Human Factors and Ergonomics Society 1*: 341-345.

Curry, R. E., & Neu, J. E. (1984, September). A model for the effectiveness of aircraft alerting and warning systems. In *Twentieth Annual Conference on Manual Control June 12-14, 1984 Ames Research* (p. 299).

Deutsch, S., & Pew, R. (2002). Modeling human error in a real-world teamwork environment. In W.D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual meeting of the Cognitive Science Society* (pp. 274–279). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Deutsch, S., & Pew, R. (2004). Examining new flight deck technology using human performance modeling. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA.

Deutsch, S., & Pew, R. (2004). Modeling the NASA SVS Part-task Scenarios in D-OMAR. BBN Report No. 8399.

Deutsch, S.E., & Pew, R.W. (2008). D-OMAR: An architecture for modeling multitask behaviors. Chapter 8 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group, pp. 183-212.

Devouassoux, Y., & Pritchett, A. (2001). Application of Kalman filtering to pilot detection of failures. In the *20th Digital Avionics Systems Conference (DASC)*, October 14-18.

Diez, M., Boehm-Davis, D.A., & Holt, R.W. (2002). Model-based predictions of interrupted checklists. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*.

Donnelly, D.M., Noyes, J.M., & Johnson, D.M. (1997). Decision making on the flight deck. *IEE Colloquium on Decision Making and Problem Solving*, pp.3/1-3/4, December 16.

Elkind, J.I., Card, S.K., Hochberg, J., & Huey, B.M. (1990). *Human Performance Models for Computer-Aided Engineering*. New York, NY: Academic Press, Inc.

Emmerson, P. (1997). Worked Example of the Oracle Target Acquisition Model. *Proceedings of the 6th NATO AGARD Meeting*. A6-1 - A6-14.

Eng, K., Lewis, R., Tollinger, I, Chu, A., Howes, A. & Vera, A. (2008) Generating automated predictions of behavior strategically adapted to specific performance objectives. *CHI 2008 Proceedings, Automatic Generation and Usability.* Montreal, Can.: Association for Computing Machinery.

Fotta, M.E., & S. Nicholson (2007). Hemets – Human error modeling for error tolerant systems. *Proceedings of the 14th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, pp 204-209.

Fowler, B. (1981). The aircraft landing test: an information processing approach to pilot selection. *Human Factors, 23*, 129-137.

Frische, F., Osterloh, J.P. & Lüdtke, A. (2010). Simulating Visual Attention Allocation of Pilots in an Advanced Cockpit Environment. *Presented at MODSIM World 2010 Conference Expo*. Pp. 713-729.

Ge., Z., Xu, H., & Liu, L. (2007). A variable strategy pilot modeling and application. *Proceedings of the 2007 IEEE International Conference on Mechatronics and Automation (ICMA)*, August 5 - 8, 2007, Harbin, China.

George, F. L. (1981). Comparison of closed loop model with flight test results. *Proceedings of the 17th Annual Conference on Manual Control*, pp. 296-301.

Gery, K., Doyal, J.,. Brett, B.,.Lebiere, C., Biefield, E., & Martin, E.A. (2003). HPMI: integrating systems engineering and human performance models. *Proceedings of the 12th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, pp 421-426.

Gil, G.H. (2010). An Accessible Cognitive Modeling Tool for Evaluation of Human-Automation Interaction in the Systems Design Process. Unpublished Doctoral Dissertation, North Carolina State University.

Gil, G., D. Kaber, S. Kim, K. Kaufmann, T. Veil, & P. Picciano (2009). Modeling pilot cognitive behavior for predicting performance and workload effects of cockpit automation. Proceedings of the 2009 International Symposium on Aviation Psychology. Dayton, OH: Wright State University, 124-129.

Glenn, F. A. III, & Doane, S.M. (1981). A Human Operator Simulator Model of the NASA Terminal Configured Vehicle (TCV). NASA Contractor Report 3421. NASA Contract NAS1-15983. Langley Research Center, Hampton, VA.

Gonzales-Calleros, J., Vanderdonckt, J., Lüdtke, A., & Osterloh, J.P. (2010). Towards model-based AHMI development. EICS '10. June 21-23, Berlin, Germany.

Gore, B.F. (2008). Human performance: Evaluating the cognitive aspects. *Handbook of Digital Human Modeling* (Ch. 32, pp. 1-18), NJ: Taylor and Francis.

Gore, B. F. (2010). The use of behavior models for predicting complex operations. *Proceedings of the Behavioral Representation in Modeling and Simulation (BRIMS) 2010*. Charleston, South Carolina.

Gore, B. F. (2010). Man-machine integration design and analysis system (MIDAS) v5: Augmentations, motivations, and directions for aeronautics applications. In P. C. Cacciabu, M. Hjalmdahl, A. Lüdtke & C. Riccioli (Eds.), *Human modelling in assisted transportation*. Heidelberg: Springer.

Gore, B.F. & Corker, K.M., (2000a). Human performance modeling: Identification of critical variables for national airspace Safety. In the Human Factors and Ergonomics Society Annual Meeting Proceedings. Santa Monica, CA: HFES.

Gore, B.F. & Corker, K.M., (2000b). Value of human performance cognitive predictions: a free flight integration application. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: HFES.

Gore, B.F. & Corker, K.M. (2002). Increasing aviation safety using human performance modeling tools: An air man–machine design and analysis system application. In M. J. Chinni (Ed.) 2002 *Military, Government and Aerospace Simulation, 34*(3), 183-188. San Diego: Society for Modeling and Simulation International.

Gore, B. F., Hooey, B. L., & Foyle, D. C. (2011, March 21-26). NASA's use of human performance models for NextGen concept development and evaluation. *In* the *20th Annual Conference on Behavioral Representation in Modeling and Simulation 2011 (BRIMS 2011)*, Sundance, UT.

Gore, B. F., Hooey, B. L., Haan, N., Bakowski, D. L., & Mahlsted, E. (2011, July 9 - July 14). A methodical approach for developing valid human performance models of flight deck operations. *In* the *Human Computer Interaction International (HCII) 2011*, Orlando, FL.

Gore, B.F., Hooey, B.L., Mahlstedt, E., & Foyle, D.C. (2013). Evaluating NextGen closely spaced parallel operations concepts with human performance models (Part 2 of 2), HCSL Technical Report (HCSL-13-02). Moffett Field, CA: NASA Ames Research Center.

Gore, B. F., Hooey, B. L., Socash, C., Haan, N., Mahlsted, E., Bakowski, D. L., Gacy, A.M., Wickens, C.D., Gosakan, M., & Foyle, D. C. (2011). Evaluating NextGen closely spaced parallel operations concepts with human performance models. HCSL Technical Report (HCSL-11-01). Moffett Field, CA: NASA Ames Research Center.

Gore, B. F., Hooey, B. L., Wickens, C.D., Socash, C., Gacy, A.M., Brehon, M, Gosakan, M., Foyle, D. C. (2013, in process). *The MIDAS workload model*. HCSL Technical Report. Moffett Field, CA: NASA Ames Research Center.

Goto, N., Chatani, K., & Fuj, S. (1995). H∞-model of the human pilot controlling unstable aircraft. *IEEE International Conference on Systems, Man and Cybernetics, Intelligent Systems for the 21st Century*, pp.2657-2662, October 22-25.

Govindaraj, T., & Mitchell, C.M. (1994). Operator Modeling in Commercial Aviation: Cognitive Models, Intelligent Displays, and Pilot's Assistants. (NASA #NCC 2-675; 90-55). Washington, DC: National Aeronautics and Space Administration.

Griffin, T.G.C., Young, M.S., & Stanton, N.A. (2010). Investigating accident causation through information network modelling. *Ergonomics, 53*(2), 198-210.

Hamilton, D. B., Bierbaum, C.R., & Fulford, L.A. (1991). Task Analysis/Workload (TAWL) (User's Guide) Version 4.0. Fort Rucker, AL.: Anacapa Sciences Inc.

Hayashi, M., Oman, C.M., & Zuschlag, M. (2003). Hidden markov models as a tool to measure pilot attention switching during simulated ils approaches. *Proceedings of the 12th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, pp 502-507.

Heiligers, M.M., Van Holten, T. & Boersema, T. (2003). On a computer-based prediction of pilot scanning workload and control workload. Paper presented at the 12th International Symposium on Aviation Psychology, Dayton, OH.

Heiligers, M.M., Van Holten, T. & Mulder, M. (2009). Predicting pilot task demand load during final approach. *The International Journal of Aviation Psychology, 19* (4), 391-416.

Hess, R. A. (1977). Prediction of pilot opinion ratings using an optimal pilot model. *Human Factors, 19*, 459-475.

Hess, R.A. (1981). Pursuit tracking and higher levels of skill development in the human pilot. *IEEE Transactions on Systems, Man and Cybernetics, 11*(4), pp.262-273, April.

Hess, R. A. (1981). An analytical approach for predicting pilot induced oscillations. *Proceedings of the 17th Annual Conference on Manual Control*, pp. 257-271.

Hess, R. A. (1987). "A Qualitative Model of Human Interaction with Complex Dynamic Systems." *IEEE Transactions on Systems, Man and Cybernetics SMC, 17*(1): 33-51.

Hess, R.A. (2009). Analytical Assessment of Performance, Handling Qualities, and Added Dynamics in Rotorcraft Flight Control. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, pp.262-271, January 2009.

Hill, R.W. Jr. (1999). Modeling perceptual attention in virtual humans. *Proceedings of the 8th Conference on Computer Generated Forces and Behavioral Representation,* Orlando, FL, May.

Hinton, J.L. (1992). The BAE SRC visual performance model - ORACLE an overview (J.S 12138). Filton, Bristol: British Aerospace PLC.

Hoh, R.H., Smith, J.C., & Hinton, D.A. (1987). The Effects of Display and Autopilot Functions on Pilot Workload for Single Pilot Instrument Flight Rule Operations. (NASA Contractor Report 4073). Washington, DC: National Aeronautics and Space Administration.

Holt, R.W., Chong, R., Hansberger, J.T. & Boehm-Davis, D.A. (2002). Modeling crew performance with ACT-R. Technical Report, December 2002. George Mason University, Psychology Department.

Holt, R.H., Chong, R., Schoppek, W., Hansberger, J.T., and Boehm-Davis, D.A. (2002). Modeling crew interaction. Workshop on ACT-R Models of Human-System Interaction.

Hooey, B. L., Gore, B. F., Wickens, C. D., Scott-Nash, S., Socash, C., Salud, E., & Foyle, D. C. (2011). Modeling Pilot Situation Awareness. *Human Modelling in Assisted Transportation*, 207-213.

Hursh, S. R., Balkin, T. J., Miller, J. C., & Eddy, D. R. (2004). The fatigue avoidance scheduling tool: Modeling to minimize the effects of fatigue on cognitive performance. *SAE transactions*, *113*(1), 111-119.

John, B. E., Patton, E. W., Gray, W. D., & Morrison, D. F. (2012, September). Tools for Predicting the Duration and Variability of Skilled Performance without Skilled Performers. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 985-989). SAGE Publications.

John, B.E., Blackmon, M.H., Polson, P.G., Fennell, K., & Teo, L. (2009). Rapid theory prototyping: An example of an aviation task. In the *HFES 53rd Annual Meeting, 53*(12), 794-798.

Jones, R. M., Laird, J. E., Nielsen, P. E., Coulter, K. J., Kenny, P., & Koss, F. V. (1999). Automated intelligent pilots for combat flight simulation. *AI magazine*, *20*(1), 27.

Jonsson, J. E., & Ricks, W. R. (1995). *Cognitive models of pilot categorization and prioritization of flight-deck information* (Vol. 3528). National Aeronautics and Space Administration, Langley Research Center.

Kaber, D.B., Alexander, A.L., Stelzer, E.M., Kim S.H., Kaufmann, K. & Hsiang, S., (2008). Perceived clutter in advanced cockpit displays: measurement and modeling with experienced pilots. *Aviat Space Environ Med, 79*: 1007 – 18.

Kaljouw, W.J., Mulder, M., & van Paassen, M.M. (2004). Multi-loop identification of pilot's use of central and peripheral visual cues. *Proceedings of the AIAA Modelling and Simulation Technologies Conference and Exhibit*. Providence, RI.

Karikawa, D., Takahashi, M., Ishibashi, A., Wakabayashi, T., & Kitamura, M. (2006). Human-machine system simulation for supporting the design and evaluation of reliable aircraft cockpit interface. *SICE-ICASE International Joint Conference*, pp.55-60, October 18-21.

Keller, J., Lebiere, C., & Shay, R. (2004). Cockpit system situational awareness modeling tool. In *Proceedings of the Human Performance, Situation Awareness and Automation Conference (HPSAA II 2004)*, Daytona Beach, FL.

Laudeman, I.V., & Palmer, E.A. (1995). Quantitative measurement of observed workload in the analysis of aircrew performance. *The International Journal of Aviation Psychology, 5(*2), 187-197.

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*. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 105-144.

Laughery, KR (1989). Micro Saint: A tool for modeling human performance in systems. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R. Sutton, & L.V. Breda (eds.) Applications of human performance models to system design (Defense research series, Vol. 2). New York City, NY: Plenum Press.

Levison, W.H. (1989). Alternative treatments of attention-sharing within the optimal control model. *IEEE International Conference on Systems, Man and Cybernetics*, pp.744-749, November 14-17.

Liu, J.Q., & Gao, Z.H. (2010). A test evaluation of a Pilot-Induced-Oscillation prediction criterion. *IEEE 2nd International Conference on Signal Processing Systems (ICSPS),* July 5-7.

Lohrenz, M.C., & Hansman,J., (2004). Investigating Issues of Display Content vs. Clutter During Air-to-Ground Targeting Missions. *In the Proceedings of the Human Factors & Ergonomics Society Annual Meeting* . Santa Monica, CA.

Lüdtke, A. & Osterloh, J-P. (2009). Simulating perceptive processes of pilots to support system design. Human Computer Interaction - INTERACT. *In the Proceedings 12th IFIP TC 13 International Conference Part I*, Uppsala, Sweden, August 24-28, 2009. (pp. 471-484).

Lüdtke, A. & Osterloh, J-P. (2010). Modeling memory effects in the operation of advanced flight management systems. *Paper presented at the 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., & Janssen, J. (2009). Capability test for a digital cognitive flight crew model. *In the Proceedings of the 3rd International Conference on Applied Human Factors and Ergonomics, AHFE 2010* (p. 1-10).

Lüdtke, A., Osterloh, J-P., Mioch, T., Rister, F., & Looije, R. (2010). Cognitive modelling of pilot errors and error recovery in flight management tasks. In the *Proceedings of* 7th IFIP WG 13.5 Working Conference, HESSD 2009, Brussels, Belgium, September 23-25, 2009, Revised Selected Papers, (pp 54-67) .

Lüdtke, A., Weber, L., Osterloh, J. P., & Wortelen, B. (2009). Modeling pilot and driver behavior for human error simulation. *Digital Human Modeling*, 403-412.

Lyall, E. A., & Cooper, B. (1992). The impact of trends in complexity in the cockpit on flying skills and aircraft operation. In the *36 th Human Factors Society Annual Meeting, Atlanta, GA* (pp. 1181-1184).

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

Martin, L., S. Verma, A. Jadhav, V. Raghavan, & S. Lozito (2003). An initial model of data link use in the cockpit. *Proceedings of the 12th International Symposium on Aviation Psychology*. (pp 769-774), Dayton, OH: The Wright State University.

McCoy, M. S. & Levary, R.R. (2000). A rule-based pilot performance model. *International Journal of Systems Science, 31*(6): 713-729.

McMillan, G.R., Beevis, D., Stein, W., Strub, M.H., Salas, E., Sutton, R., & Reynolds, K.C. (1991). A Directory of Human Performance Models (AC/243 (Panel 8)TR/1). Brussels: NATO Headquarters.

McMillan, G.R., Beevis, D., Salas, E., Strub, M.H., Sutton, R., & Breda, L.V. (1989). Applications of human performance models to system design (Defense research series, Vol. 2). New York City, NY: Plenum Press.

McNally, B.H. (2005). An approach to human behavior modeling in an air force simulation. *Proceedings of the 2005 Winter Simulation Conference*, pp.5 pp., December.

Milgram, P., van der Wijngaart R., Veerbeek, H., Fokkerweg, A., & Bleeker, O. (1984). Multi-crew model analytic assessment landing performance and decision making demands. *Proceedings of the 20th Annual Conference on Manual Control, volume 2*, (pp. 374-396).

Miller, C. A., (1998). *Case Studies Involving W/Index,* Honeywell Technology Center.

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.

Mioch, T., Mistrzyk, T., & Rister, F. (2010). Procedure Design and Validation by Cognitive Task Model Simulations. In *Proceedings of the 19th Conference on Behavior Representation in Modeling and Simulation. Charleston, SC, USA* (pp. 232-239).

Mioch, T., Osterloh, J-P, & Javaux, D. (2010). Selecting human error types for cognitive modelling and simulation. *Paper presented at the Human Modelling of Assisted Technologies Workshop*, Begirate, Italy.

Mohlenbrink, C., Lenz, H., Korn, B. (2010). An overview of eye-movement analyses measures for validating a cognitive pilot model. *Paper presented at the Human Computer Interaction Aero Conference 2010*, Cape Canaveral, FL.

Mulgund, S., Rinkus, G., Illgen, C., & Zacharias, G. (1997). Situation awareness modeling and pilot state estimation for tactical cockpit interfaces. HCI International Conference.

Muraoka, K. & Tsuda, H. (2006). Flight Crew Task Reconstruction for Flight Data Analysis Program. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50(*11): 1194-1198.

Muraoka, K., Verma, S., Jadhav, A., Corker, K. M., & Gore, B. F. (2004). *Human Performance Modeling of Synthetic Vision System Use*. Technical Report, San Jose State University, San Jose, CA.

Nelson, W.R., (1988). Functional Models of Complex Human Performance: Application to the Assessment of Pilot Performance. *In the Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: HFES.

Nikolic, M.I., & Sarter, N.B. (2003). Towards a model of error management on highly automated glass cockpit aircraft. *Proceedings of the 12th International Symposium on Aviation Psychology* (pp 882-887). Dayton, OH: Wright State University.

Osterloh, J-P, & Lüdtke, A. (2008). Analyzing the ergonomics of aircraft cockpits using cognitive models. *Proceedings of the 2nd Applied Human Factors and Ergonomics*. 1-10.

Parks, D. & Boucek, G. Workload prediction, diagnosis and continuing challenges. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R. Sutton, & L.V. Breda (1989). *Applications of human performance models to system design (Defense research series, Vol. 2*). New York City, NY: Plenum Press.

Pisanich, G. M., & Corker, K. M. (1995, April). A predictive model of flight crew performance in automated air traffic control and flight management operations. In *Proceedings of the 8th international symposium on aviation psychology* (pp. 335-340).

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

Prasad, S.N. & Schmidt, D.K. (1980). Multi-axis tracking via an optimal control pilot model. *Proceedings of the 16th Annual Conference on Manual Control*, p. 115.

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.

Rao, A. S., Morley, D., Sekvestrel, M., & Murray, G. (1992, November). Representation, selection, and execution of team tactics in air combat modelling. In *Proc. 5th Australian Joint Conference on AI* (pp. 185-190)

Remington, R., Matessa, M., Freed, M., & Lee, S. (2003). Using Apex/CPM-GOMS to develop human-like software agents. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems.* Melbourne: ACM Press.

Rickard, W. W., & Levison, W. H. (1981). Further tests of a model-based scheme for predicting pilot opinion ratings for large commercial transports. *Proceedings of the 17th Annual Conference on Manual Control,* pp. 247-256.

Riley, V., Lyall, E., Cooper, B., & Wiener, E. (1991) Analytic methods for flight-deck automation design and evaluation. Phase 1 report: flight crew workload prediction. FAA Contract DTFA01-91-C-00039. Minneapolis Minn: Honeywell Technical Center.

Ross, L. E., & Mundt, J. C. (1988). Multiattribute modeling analysis of the effects of a low blood alcohol level on pilot performance. *Human Factors, 30*, 293-304.

Rouse, W.B., Hammer, J.M., Mitchell, C.M., Morris, N.M., Lewis, C.M., & Yoon, W.C. (1985). Pilot interaction with automated airborne decision making systems. NASA Grant NAG 2-123. Washington, DC: National Aeronautics and Space Administration.

Rushby, J. (2002). Using model checking to help discover mode confusions and other automation surprises. *Reliability Engineering & System Safety, 75* (2), Feb 2002, pp 167-177.

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., Young, M.S., Harris, D., Demagalski, J., Marshall, A., Waldmann, T., & Dekker, S. (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*,.(pp 567-571).

Sarno, K., & Wickens, C. (1995). The role of multiple resources in predicting time-sharing efficiency: An evaluation of three workload models in a multiple task setting. *International Journal of Aviation Psychology, 5*(1), 107-130.

Schmidt, D. K. (1981). On the use of the ocm's quadratic objective function as a pilot rating metric. *Proceedings of the 17th Annual Conference on Manual Control*, pp. 306-314.

Schoelles. M.J., & Gray, W.D. (2011). Cognitive modeling as a tool for improving runway safety. The *Proceedings of the 16th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, 541-546.

Schoppek, W., & Boehm-Davis, D. A. (2004). Opportunities and challenges of modeling user behavior in complex real world tasks. *MMI interaktiv*, *7*, 47-60.

Schulte, A., & R. Onken (1995). Modeling of pilot's visual behavior for low-level flight. *Proceedings of Synthetic Vision for Vehicle Guidance and Control AeroSense '95*, Orlando, FL (17-21 April 1995).

Schurr, N. (2011). ALARMS: Alerting and reasoning management system. *Presentation delivered to the 2011 NASA Aviation Safety Technical Meetin*g, St. Louis, MO.

Sebok, A., Wickens, C., Sarter, N., Quesada, S., Socash, C., Anthony, B. (2012). The Automation design advisor tool (ADAT): Development and validation of a model-based tool to support flight deck automation design for nextgen operations. *Human Factors and Ergonomics in Manufacturing and Service Industries, 22*(5), 378-394.

See, J.E., & Vidulich, M.A. (1998). Computer modeling of operator mental workload and situational awareness in simulated air-to-ground combat: An assessment of predictive validity. *The International Journal of Aviation Psychology, 8*(4), 351-375.

Shively, R. J., Brickner, M., & Silbiger, J. (1997). A computational model of situational awareness instantiated in MIDAS. .*Proceedings of the Ninth International Symposium on Aviation Psychology*, Dayton, OH: Wright State University.

Siesfeld, A., Curley, R., & Calfee, I. (1984). Communication on the flight deck. *Proceedings of the 20th Annual Conference on Manual Control*, Volume 2, pp.265-275.

Smith, S.C., Govindaraj, T., & Mitchell, C.M., (1990). Operator modeling in civil aviation. *IEEE International Conference on Systems, Man and Cybernetics*, pp.512-514, November 4-7

Sorensen, J., & Goka, T. (1984). Predictions of cockpit simulator experimental outcome using system models, pp 269-290.

Stanton, N.A., Salmon, P., Harris, D., Demagalski, J., Marshall, A., Waldmann, T., & Dekker, S. (2003). Predicting pilot error: Assessing the performance of SHERPA. *Proceedings of the HCII Conference*, pp 587-591.

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

Stokes, A.F. & Raby, M., (1989). Stress and cognitive performance in trainee pilots. *In the Proceedings of the Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: Human Factors Society.

Stone, G., Culick, R. & Gabriel, R. (1987) Use of task timeline analysis to assess crew workload. In A. Roscoe (Ed.), *The practical assessment of pilot workload.* NATO AGARDograph #282.

Stroeve, S.H., & Blom, H.A.P. (2005). Human performance modeling for accident risk assessment of active runway crossing operation. *Proceedings of the 2005 International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, 725-730.

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.

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. In the *9th USA/Europe ATM R&D seminar*.

Stütz, P., & Onken, R. (1997). Adaptive Pilot Modeling within Cockpit Crew Assistance. *Advances in human factors/ergonomics*, 733-736.

Svensson, E., Rencrantz, C., Lindoff, J., Berggren, P., & Norlander, A. (2006, October). Dynamic Measures for Performance Assessment in Complex Environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 24, pp. 2585-2589). SAGE Publications.

Svensson, E.A.I., & Wilson, G.F. (2002). Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. *The International Journal of Aviation Psychology*, *12*(1), 95-110.

Swauger, S. (2003). How good pilots make bad decisions: A model for understanding and teaching failure management to pilots. *Proceedings of the 12th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, pp. 1137-1142.

Thomas, M. J. W. (2004). Predictors of threat and error management: Identification or core nontechnical skills and implications for training systems design." *International Journal of Aviation Psychology, 14*(2): 207-231.

Tidhar, G., Heinze, C., & Selvestrel, M. (1998). Flying together: Modelling air mission teams. *Applied Intelligence*, *8*(3), 195-218.

Tidhar, G., Selvestrel, M., & Heinze, C. (1995, April). Modelling teams and team tactics in whole air mission modelling. In *Proceedings of the Eighth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE'95)* (pp. 373-381).

Uijtde Haag, M., Duan, P., Schnell, T., Cover, M., Anderson, N., Snow, M., Etherington, T., Rademaker, R., & Theunissen, E. (2011). Hazard and Integrity Monitoring and Integrated Alerting and Notification Methods. Presentation delivered to the 2011 NASA Aviation Safety Technical Meeting in St. Louis, MO.

Van Dongen, H.P.A. (2004). Comparison of mathematical model predictions to experimental data of fatigue and performance. *Aviat Space Environ Med*; 75 (Suppl 1): A15-A36

Verfurth, S.C. Govindaraj, T., & Mitchell, C.M. (1991). OFMspert for the 727: an investigation into intent inferencing on the flight deck. *IEEE International Conference on Systems, Man, and Cybernetics, Decision Aiding for Complex Systems*, pp.1311-1316, October 13-16.

Verma, S., Corker, K. (2002). Introduction of context in a human performance model to predict performance for new air traffic management initiatives. *Proceedings of the Advanced Simulation Technologies Conference 2002*, San Diego, CA.

Verma, S.A., Corker, K. & Jadhav, A., (2003). An approach to modeling error in Air-MIDAS using contextual control model. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA: Human Factors Society.

Walden, R.S., & Rouse, W.B. (1978). A queueing model of pilot decision making in a multitask flight management situation. *IEEE Transactions on Systems, Man and Cybernetics*. pp.867-875, December 1978.

Washizu, K., Tanaka, K., & Osawa, T. (1980). An experimental study of human pilot's scabning behavior. *Proceedings of the 16th Annual Conference on Manual Control*, pp. 138-144.

Wewerinke, P.R., (1980). The effect of visual information on the manual approach and landing. *Proceedings of the 16th Annual Conference on Manual Control*, pp. 58-74.

Wickens, C.D., Harwood, K., Segal, L., Tkalcevic, I., & Sherman, B. (1988). TASKILLAN: A simulation to predict the validity of multiple resource models of aviation workload. *Proceedings of the 32nd Meeting of the Human Factors Society* (pp. 168-172). Santa Monica, CA: Human Factors Society.

Wickens, C.D., Bagnall, T., Gosakan, M., & Walters, B. (2011). A Cognitive Model of the Control of Unmanned Aerial Vehicles. The *Proceedings of the 16th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University. 535-540.

Wickens, C.D. (1989) Models of multi-task situations. In McMillan, G.R., Beevis, D., Salas, E., Strub, M.H., Sutton, R., Breda, L.V. (1989). *Applications of human performance models to system design (Defense research series, Vol. 2).* New York City, NY: Plenum.

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

Wickens, C. D., Hooey, B. L., Gore, B. F., Sebok, A., & Koenicke, C. S. (2009). Identifying black swans in nextgen: predicting human performance in off-nominal conditions. *Human Factors, 51*, 638-651.

Wickens, C.D., Larish, I. & Contoror, A. (1989). Predictive Performance Models and Multiple Task Performance. Proceedings of the Human Factors Society 33rd Annual Meeting, pp 96-100.

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. Chapter 9 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 213-242.

Wickens, C. D., Sandry, D. L., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *25*(2), 227-248.

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.

Xiaoru, W., Damin, Z., & Hengyang, W. (2009). Pilot attention allocation model in complicated human-machine interface. *International Conference on Biomedical Engineering and Informatics* (BMEI), pp.1-5, October 17-19.

Xiaoru, W., Hengyang, W., & Damin, Z. (2010). Study on pilot attention allocation model based on fuzzy theory. Sixth International Conference on Natural Computation (ICNC), August 10-12, pp. 2035-2039.

Zaal, P. M. T., Pool, D. M., Chu, Q. P., Van Paassen, M. M., Mulder, M., & Mulder, J. A. Delft University of Technology, 2600 GB Delft, The Netherlands.

Zacharias, G. L., Miao, A. X., Illgen, C., Yara, J. M., & Siouris, G. M. (1996). SAMPLE: Situation awareness model for pilot in-the-loop evaluation. *Final Report R*, *95192*.

Zacharias, G., Warren, R., & Riccio, G. (1986). Modeling the pilot's use of flight simulator visual cues in a terrain-following task. *In the 22nd Annual Conference on Manual Control*, pp 81-82, Belton Inn, Dayton, Ohio, July 15th-16th, 1986.

Zuschlag, M., (2004). Quantification of visual cluttering using a computational model of human perception: An application for head-up displays. In *Proceedings of the Human Performance, Situation Awareness and Automation Conference (HPSAA II 2004),* Daytona Beach, FL.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED *(From – To)*
30-04-2013	Technical Memorandum	

4. TITLE AND SUBTITLE

Modeling and Evaluating Pilot Performance in NextGen: Review of and Recommendations Regarding Pilot Modeling Efforts, Architectures, and Validation Studies

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Christopher Wickens, Angelia Sebok, John Keller, Steve Peters, Ronald Small, Shaun Hutchins, Liana Algarin, Brian F. Gore, Becky L. Hooey, David C. Foyle

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

DTFAWA-10-X-80005

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES(ES)

NASA Ames Research Center
Moffett Field, California 94035-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

TH-094

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING REPORT NUMBER

NASA/TM–2013-216504

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified—Unlimited
Subject Category: 1 Availability: NASA CASI (301) 621-0390 Distribution: Nonstandard

13. SUPPLEMENTARY NOTES

Point of Contact: David Foyle, NASA Ames Research Center, 262-2, Moffett Field, CA; (650) 604-3053

14. ABSTRACT

NextGen operations are associated with a variety of changes to the national airspace system (NAS) including changes to the allocation of roles and responsibilities among operators and automation, the use of new technologies and automation, additional information presented on the flight deck, and the entire concept of operations (ConOps). In the transition to NextGen airspace, aviation and air operations designers need to consider the implications of design or system changes on human performance and the potential for error. To ensure continued safety of the NAS, it will be necessary for researchers to evaluate design concepts and potential NextGen scenarios well before implementation. One approach for such evaluations is through human performance modeling. Human performance models (HPMs) offer advantages over empirical, human-in-the-loop testing in that they allow detailed analyses of systems that have not yet been built; offer flexibility for extensive data collection; and they don't require experimental participants. HPMs differ in their ability to predict performance and safety with Nexten procedures, equipment and ConOps. Our research objectives were to support the FAA in identifying HPMs appropriate for predicting pilot performance in NextGen operations, provide guidance on how to evaluate the quality of different models, and to identify gaps inpilot performance modeling research that could guide future research. This research is intended to help the FAA evaluate pilot modeling efforts and select the appropriate tools for future modeling efforts to predict pilot performance in NextGen operations.

15. SUBJECT TERMS Pilot human performance models; NextGen; Human performance

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Help Desk at email: help@sti.nasa.gov
U	U	U	UU	147	19b. TELEPHONE NUMBER (Include area code) STI Help Desk at: (301) 621-0390

Modeling Effort	*Model or Architecture*	*Deep dive in which it was evaluated*	*Verification*				*Validation*			
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Blom, Corker, Stroeve & van der Park (2003)	Air MIDAS	E	X							
Boehm-Davis, Holt, Chong & Hansberger (2004)	ACT-R	R&R				X				
Boehm-Davis, Holt, Diez & Hansberger (2002)	ACT-R	A, SA					X			
Burdick & Schively (2000) / Shively, Brickner & Silbiger (1997)	MIDAS	SA								X
Byrne, Kirlik & Fleetwood (2008)	ACT-R	E					X			
Carlin, Alexander & Schurr (2011)	ALARMS	A	X							
Corker & Pisanich (1998)	Air MIDAS	R&R					X			
Corker, Muraoka, Verma, Jadhav & Gore (2008)	Air MIDAS	E					X			
Deutsch & Pew (2004)	D-OMAR	A					X			
Deutsch & Pew (2008)	D-OMAR	E, WL, MT					X			
Donnelly, Noyes & Johnson (1997)	IDM	SA	X							
Fotta, Nicholson & Byrne (2007)	ACT-R	E	X							
Gil, Kaber, Kim, Kaufmann, Veil & Picciano (2009)	E-GOMSL	A, WL								X
Gonzales-Calleros, Vanderdonckt, Lüdtke & Osterloh (2010)	Usability Advisor	A	--							
Gore & Corker (2000a)	Air MIDAS	WL			X					
Gore & Corker (2000b)	Air MIDAS	WL			X					
Gore (2008)	MIDAS	WL	--							
Gore, Hooey, Mahlstedt & Foyle (2013)	MIDAS	R&R				X				

Modeling Effort	*Model or Architecture*	*Deep dive in which it was evaluated*	*Verification*				*Validation*			
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Gore, Hooey, Socash, Haan, Mahlsted, Bakowski, Gacy, Wickens, Gosakan & Foyle (2011)	MIDAS	WL								X
Hooey, Gore, Wickens, Salud, Scott-Nash, Socash & Foyle (2010)	MIDAS	SA, R&R				X				
John, Blackmon, Polson, Fennell & Teo (2009)	ACT-R	A				X				
Karikawa, Takahashi, Ishibashi, Wakabayashi & Kitamura (2006)	PCS	E, SA					X			
Keller, Lebiere & Shay (2004)	ACT-R	SA			X					
Laudeman & Palmer (1995)	TLAP	WL, MT								X
Lebiere, Archer, Best & Schunk (2008)	ACT-R	E, WL					X			
Lüdtke & Osterloh (2010)	CASCaS	A					X			
Lüdtke, Osterloh & Frische (2012)	CASCaS	A							X	
Lüdtke, Osterloh, Mioch, Rister & Looije (2009)	CASCaS	E, A					X			
Lyall & Cooper (1992)	MRT	R&R, WL, MT		X						
Manton & Hughes (1990)	MRT	A, WL				X				
McNally (2005) / Zacharias, Miao et al (1996)	SOAR	SA		X						
Miller (1998)	MRT	WL		X						
Miller (2001)	ASHRAM	E	--							
Muraoka & Tsuda (2006)	OPSAMS	WL	X							
Nikolic & Sarter (2003)	(ASRS data)	E								X
Parks & Bouceck (1989)	TLAP	WL								X
Pisanich & Corker (1995)	Air MIDAS	A, R&R					X			
Polson & Javaux (2001)	GOMS	A, MT		X						

Modeling Effort	*Model or Architecture*	*Deep dive in which it was evaluated*	*Verification*				*Validation*			
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Raeth & Reising (1997)	(trust / workload allocation)	A		X						
Rickard & Levison (1981)	OCM	WL								X
Riley, Lyall, Cooper & Wiener (1991)	MRT	MT, WL								X
Salmon, Stanton, Young, Harris, Demagalski, Marshall, Waldman & Dekker (2002)	SHERPA	E					X			
Salmon, Stanton, Young, Harris, Demagalski, Marshall, Waldman & Dekker (2003)	SHERPA	E					X			
Sarno & Wickens (1995)	MRT	WL						X		
Schoelles & Gray (2011)	ACT-R	MT		X						
Schoppek & Boehm-Davis (2004)	ACT-R	A, MT					X			
Schurr (2011)	ALARMS	WL	X							
Sebok, Wickens, Sarter, Quesada, Socash & Anthony (2012)	SEEV, N-SEEV, ADAT	A, WL								X
See & Vidulich (1998)	Microsaint	SA, WL							X	
Stanton, Salmon, Harris, Demagalski, Marshall, Waldman & Dekker (2003)	SHERPA	E					X			
Steelman-Allen, McCarley & Wickens (2011)	N-SEEV	WL								X
Stone, Culick & Gabriel (1987)	TLAP	R&R, WL				X				
Stroeve & Blom (2005)	TOPAZ	E		X						
Stroeve, Blom & Bakker (2009)	TOPAZ	E, SA, R&R			X					
Stroeve, Blom & Bakker (2011)	TOPAZ	E, SA			X					

Modeling Effort	*Model or Architecture*	*Deep dive in which it was evaluated*	*Verification*				*Validation*			
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Svensson & Wilson (2002)	(regression)	SA					X			
Tidhar, Heinze & Selvestrel (1998)	SWARMM	R&R	X							
Uijtde Haag, Duan, Schnell, Cover, Anderson, Snow, Etherton, Rademaker & Theunissen (2011)	SOAR	A	--							
Walden & Rouse (1978)	(queueing)	MT					X			
Wickens, Bagnall, Gosakan & Walters (2011)	N-SEEV	MT		X						
Wickens, Goh, Helleberg, Horrey & Talleur (2003)	SEEV	MT, SA							X	
Wickens, Harwood, Segal, Tkalcevic & Sherman (1988)	TLAP, MRT	MT, WL							X	
Wickens, Hooey, Gore, Sebok & Koenecke (2009)	N-SEEV	SA							X	
Wickens, Larish & Contoror (1989)	TLAP, MRT	WL							X	
Wickens, McCarley, Alexander, Thomas, Ambinder & Zheng (2008)	A/SA	E, MT, SA					X			
Wickens, Sebok, Kamienski & Bagnall (2007)	A/SA, SEEV	SA				X				

Notes: A-Automation interaction, E-Error, MT – Multi-task, SA – Situation Awareness, WL – Workload, R&R – Roles and Responsibility

4. Model Architectures

4.1 Overview

Model architectures are defined differently than individual modeling efforts. A model architecture is a framework for creating any number of more specific models to predict hu performance. The architecture can be a framework for representing cognition or attention ACT-R or SEEV) that eventually needs to be implemented within a programming languag an architecture can be a specific modeling tool that allows users to create models (e.g., IMPRINT).

Across our reviews of modeling efforts, through all five deep dives, we have identified sev architectures that were used in multiple modeling efforts, and were often used to model m aspects of pilot performance (e.g., situation awareness, workload). We describe these architectures below. While these descriptions may be partially redundant with the specific effort applications discussed in Section 3, our emphasis here is somewhat different: how d model work, how might it contribute to several different modeling aspects, and how usabl available is it?

Table 4.1 provides a list of key architectures and the aspects of pilot performance that they modeled. These data were compiled from the spreadsheet developed as part of this researc Each modeling effort that could be associated with a specific architecture (e.g., ACT-R, M was identified. The specific aspects of pilot performance that were modeled in these effor also identified and indicated with *'s in the table. Finally, the extent to which the architec had been validated was calculated by taking the total number of modeling efforts (again, i spreadsheet) for a given architecture and dividing that into the total number of validated ef for that architecture.

Table 4.1: Pilot modeling aspects and validation efforts addressed by different major architectures.

Pilot Aspect:	Human-Automation Interaction	Communication	Decision making	Error	Manual Control	Multi-Tasking	Procedures	Roles & Responsibilities	Situation Awareness	Spatial Disorientation	Visual scanning	Workload	Percent of efforts that included at least a partial *Validation*
Architecture													
ACT-R	*	*	*	*	--	*	*	*	*	--	*	*	6/15; 40%
Air MIDAS	*	*	*	*	--	--	*	--	--	--	*	*	6/11; 55%
CASCaS	*	--	--	*	--	*	*	--	--	--	*	--	4/10; 40%
MIDAS	--	*	*	*	--	*	*	*	*	--	*	*	6/11; 55%
MRT	*	--	--	--	--	*	--	*	--	--	--	*	3/8; 38%
OCM	--	--	--	--	*	--	--	--	--	--	*	*	7/9; 78%
SEEV, N-SEEV, A/SA	*	--	*	--	--	*	--	--	*	--	*	--	5/7; 71%
TLAP	--	--	--	--	--	*	--	--	--	--	--	*	4/5; 80%
TOPAZ	--	*	*	*	--	--	*	--	*	--	--	--	1/6; 17%

- MRT – includes W/Index

4.2 Adaptive Control of Thought – Rational (ACT-R)

4.2.1 What Is It?

The Adaptive Control of Thought – Rational (ACT-R) is a unified theory of cognition that integrates theories of attention, cognition, and motor actions. It was developed by Carnegie Mellon University for the Office of Naval Research in 1993, and has been in use and updated multiple times since then. ACT-R is a theory of cognition that has been represented as executable software that is programmed in LISP.

ACT-R provides a framework for representing human cognition. It models cognition through "production rules" or goal-directed behavior that is implemented through a series of "if-then" rules. It includes perceptual inputs and motor outputs.

ACT-R's main components are modules, buffers, and the pattern matcher. ACT-R uses perceptual-motor modules (visual and manual modules) to simulate interaction with the physical environment. Memory modules simulate declarative and procedural memory and allow the modeler to represent an operator accessing different types of information from long-term memory. Declarative memory consists of simple facts. Procedural memory consists of representations for how to perform different tasks.

Buffers are interfaces for each of the ACT-R modules, except the procedural memory module. The contents of each buffer at a given time represent the state of ACT-R at that moment. The pattern matcher uses the buffer contents to identify a relevant schema to select goals and behavioral rules. Cognition is modeled as a succession of these changes as instructed by the pattern matcher.

ACT-R does not have a goal hierarchy, as a pre-established hierarchy would suggest perfect memory on the part of the modeled operator. With the pattern matcher, whatever situation provides the closest match is what changes the state of the system.

Another important aspect of ACT-R is that it models the actual time required for cognitive steps (e.g., retrieving and 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 times, provided as default parameters in ACT-R, are based on psychological theories or empirical research (e.g., 50 msec to retrieve information; Fitts' law for scanning a display or selecting a control). In addition to timing aspects, ACT-R models errors of omission, where a retrieval from memory fails, and errors of commission, where an error occurs due to imperfect matching.

ACT-R 6 version 1.4 [r1261] (the latest version as of August 9, 2012) is available on the website: http://act-r.psy.cmu.edu/actr6/

Inputs: Data from psychology experiments, general assumptions about human cognition, assumptions about a particular domain

Outputs: overt behavior, time to perform the task, accuracy in the task, and whether or not the task was performed

4.2.2 What Has It Been Applied To?

ACT-R has been used in aviation, and in many other domains. These include human-computer interaction, department of defense research, education (in cognitive tutoring systems), and neuropsychology. It has been used to predict the time to complete task sequences (Fleetwood, Lebiere, Archer, Mui, & Gosakan, 2006), if operator error occurs (John et al., 2009), and the type of error (Boehm-Davis et al., 2002).

ACT-R models have been used in more than 700 different scientific publications.
A long list of references is available: http://act-r.psy.cmu.edu/publications/index.php.

4.2.3 Where Can You Get It?

ACT-R is an open-source architecture, available online at:
http://act-r.psy.cmu.edu/

4.2.4 How Usable Is It?

ACT-R must be implemented using the LISP programming language. The CMU ACT-R website offers numerous resources for helping potential ACT-R modelers in learning how to use the tool. Tutorials and manuals are available, free of charge, online http://act-r.psy.cmu.edu/actr6/. In addition, CMU hosts a summer school and workshops for learning to use ACT-R: http://act-r.psy.cmu.edu/workshops/ Further, CMU provides a contact person for support db30@andrew.cmu.edu

4.2.5 How Extensively Validated Is It?

ACT-R is extensively validated. Validation efforts are used to update and refine the cognitive theory behind the model.

4.3 Attention – Situation Awareness (A/SA)

4.3.1 What Is It?

The A/SA (Attention – Situation Awareness) modeling architecture was developed by researchers at the University of Illinois for NASA (Wickens, McCarley et al., 2008). It predicts operator situation awareness by using an attention model and a belief-updating module. The underlying attention model is SEEV (salience, effort, expectancy, and value).

SEEV predicts operator visual scanning, and says that scanning is driven by two bottom-up factors (salience and effort) and two top-down factors (expectancy and value). The salience, or obviousness, of a visual indication increases the likelihood that a cue will draw the operator's attention. Expectancy, the expectation that the information is changing frequently (events occurring rapidly) and therefore needs to be sampled often, also affects how often the operator looks at a display. Value, or the importance of information, increases the likelihood that the operator will view the information at the location of the valued commodity. Effort, the difficulty associated with moving attention to a display – either due to distance from the current visual focal point, or the need to navigate among pages in a multifunctional display – is the only factor that decreases the

likelihood of an operator viewing a display more often. Experienced operator scanning is driven relatively more by top-down factors (primarily by expectations and importance of information). The model runs as a discrete event simulation of SEEV, with probabilistic movement to attentional locations, to a degree proportional to the overall contributions of the four components.

Situation awareness is modeled according to two stages of the Endsley (1988) three-stage model, where stage 1 represents perception and stage 2 is comprehension (See Section 3.3 for elaboration of these concepts). For an operator to have awareness, s/he must perceive or attend to the relevant display. SEEV predicts whether or not the operator perceives the data. Once stage 1 awareness is attained, the belief-updating module in the A/SA architecture makes further predictions. A/SA models situation awareness on a scale of 0 to 1, where 1 is perfect awareness. Each display is associated with a particular type of awareness that the operator needs. When the display is viewed, awareness for the parameters on that display increases to 1. Over time, awareness decreases by a postulated decay function. For accurate information, the awareness decreases after a minute, but if distracting information is viewed, awareness decreases much more rapidly. The stage 1 SA of individual parameters is aggregated, and when sufficient parameter awareness exists, the operator is said to have achieved higher stages of SA. The overall awareness, in turn, drives scanning behavior. If the modeled operator has good SA, scanning is driven primarily by expectancy and value. Conversely, if the modeled operator has poor SA, sub-optimal scanning results. In the application reported in Wickens, McCarley et al. (2008), SA affects operator decision making, with better SA resulting in better decisions.

Visual scanning may be driven by habit, as in SEEV; but also by salient time stamped **events** that capture attention, such as the onset of a single warning indicator. The N-SEEV (Noticing-SEEV) model is an extension of SEEV (also developed by University of Illinois for NASA), and predicts operator noticing of a change in a visual field (McCarley et al., 2009; Sebok et al., 2006; Steelman-Allen & Wickens, 2011; Wickens, Hooey et al., 2009). The N-SEEV model uses SEEV to predict scanning, with the difference that a change occurs at some point in a simulation. The change is associated with a pre-defined location and salience in the visual field. Operator scanning continues, and if the operator views the changed display within a given time frame (typically within 10 seconds of the change), the operator is said to have noticed the change. If the operator does not view the change within that time frame, N-SEEV predicts that change blindness will occur, and the operator – even if s/he views the display – will not notice the difference. N-SEEV is typically run multiple times to generate a distribution of noticing times. The probability that the operator will notice the change is calculated based on the probability of runs that had fixations landing on the display where the relevant event occurred before the (e.g., 10 second) cutoff. The predicted time required for the operator to notice the change is calculated as the average noticing time for all of the "noticed" changes.

4.3.2 What Has It Been Applied To?

- SEEV: Aviation, Driving, Nursing
- A/SA: Aviation
- N-SEEV: Aviation

4.3.3 Where Can You Get It?

The A/SA, SEEV, and N-SEEV architectures are not software tools, nor are they particularly complex frameworks (like, e.g., ACT-R). They are practical models that can easily be implemented in a variety of programming languages. It is suggested that the models are developed with SME input, to identify the salience, effort, expectancy and value parameters for the visual displays in different operating conditions.

4.3.4 How Usable Is It?

The A/SA, SEEV, and N-SEEV architectures are easy to apply. No online support is available for these architectures but a wide variety of publications describes their implementation. The SEEV model can be implemented with simple analytic equations, although SMEs are required to estimate parameters of expectancy and value.

4.3.5 How Extensively Validated Is It?

A/SA, SEEV, and N-SEEV have all been validated against aviation human performance data.

4.4 The Cognitive Architecture for Safety Critical Task Simulation (CASCaS)

4.4.1 What Is It?

CASCaS (Cognitive Architecture for Safety Critical Task Simulation) was developed as part of a European Union project, HUMAN: Model-Based Analysis of Human Errors During Aircraft Cockpit System Design[1]. The intent for CASCaS was to address specifically the cognitive processes that are relevant for safety critical system design, particularly in aviation. As Lüdtke, Osterloh, and Frische (2012, p.1) state: "The innovative aspect of CASCaS is the prediction of human errors resulting from an interaction of (1) learned mental models (Routine Learning/Learned Carelessness), (2) actual limited cognitive performance and (3) safety nets in aircraft cockpit design (e.g. flashing indication, alerts and crew interaction)."

The HUMAN project included an effort to create an executable flight crew model that incorporated cognitive error producing mechanisms. This model was created using CASCaS, and it interacted with a flight simulator and a model of a cockpit interface. The model included routine learning, the learning of shortcuts, and attention allocation (Lüdtke, Osterloh, Mioch, Rister, & Looije, 2009).

The CASCaS model has been used to predict visual scanning behavior, dwell times in areas of interest (AOIs), and the time required to notice specific visual indications in the cruise and approach phases of flight (Lüdtke, Osterloh, & Frische, 2012). CASCaS, like ACT-R, is a cognitive architecture, which provides a structure and set of if-then rules for simulating human cognition.

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. These

[1] This project was performed during 1 March 2008 – 28 Feb 2011. More information is available at:
EU website: http://ec.europa.eu/research/transport/projects/items/human_en.htm
HUMAN Aero site: http://www.human.aero/

correspond, respectively, to the skill-based, rule-based, and knowledge-based levels of behavior defined by Rasmussen (1983). New tasks require significant effort on the part of an operator, and are performed at the cognitive level. These are modeled by the operator having a high-level goal, and selection of a plan by which to meet that goal. When operators are somewhat familiar with a task, they are working on the associative level. These types of tasks are modeled by selection of a set of rules. Finally, frequently performed tasks are autonomous, and performed without conscious thought. These include maneuvering the aircraft (Lüdtke, Osterloh, Mioch, Rister, & Looije, 2009).

In solving problems, humans typically apply the easiest solution – autonomous, if possible; associative, if they can identify a relevant pattern for the situation; and cognitive if absolutely necessary due to lack of established rules or safety-critical concerns. The Lüdtke, Osterloh, et al., (2009) team focus on error mechanisms at the associative and cognitive levels.

Knowledge in the associative and cognitive layers is stored in the memory component. Short term memory includes data that have been perceived by the modeled operator from the environment or derived from rules. Long term memory stores flight procedures as "if then" rules. When the "if" condition is met, the "then" condition is triggered. Figure 4.1 below shows an overview of the CASCaS components.

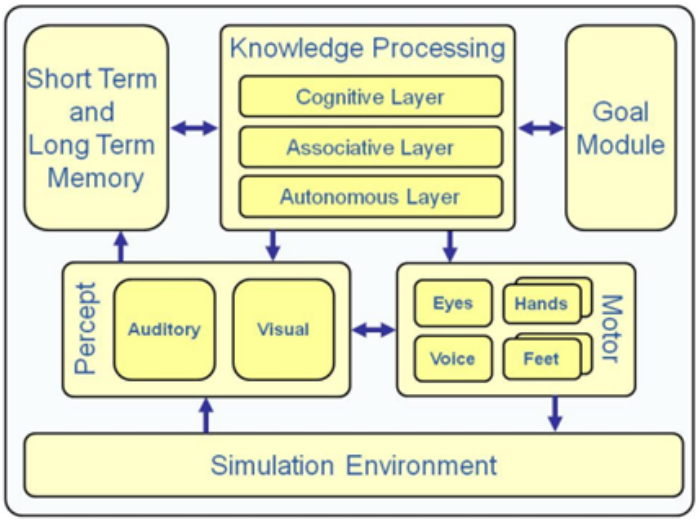

Figure 4.1. Overview of CASCaS Components
From http://www.isi-padas.eu/sites/default/files/ISi-PADAS_Newsletter_5.pdf

In another CASCaS effort (Lüdtke, Osterloh, Mioch, Rister, & Looije, 2009), the CASCaS model predicts cognitive errors such as Learned Carelessness and Cognitive Lockup (See 3.1 and 3.4 for descriptions of these applications).

4.4.2 What Has It Been Applied To?

CASCaS has been applied to a variety of aviation tasks, but primarily focusing on pilot error mechanisms. It has also been applied to modeling vehicle driver performance (http://www.isi-padas.eu/sites/default/files/ISi-PADAS_Newsletter_5.pdf).

4.4.3 Where Can You Get It?

It does not appear that CASCaS is available, neither as open source product nor for purchase. For more information, the reader should contact Dr. Andreas Lüdtke (Lüdtke@offis.de).

4.4.4 How Usable Is It?

This is not relevant, since the architecture is not available.

4.4.5 How Extensively Validated Is It?

CASCaS has been quite extensively validated as part of the EU project that funded it. The effort included many modeling phases as well as empirical, pilot in the loop studies. Validation efforts focused on particular aspects of performance (e.g., when an error occurred in seeking additional, safety-critical information).

4.5 The Man-machine Integration Design and Analysis System (MIDAS)

4.5.1 What Is It?

The Man-machine Integration Design and Analysis System (MIDAS; http://humansystems.arc.nasa.gov/groups/midas/) is an established HPM that predicts human-system performance under nominal and off-nominal conditions (Gore, 2010). MIDAS is a dynamic, integrated human performance model environment that facilitates the design, visualization, and computational evaluation of complex man-machine system concepts in simulated operational environments. MIDAS symbolically represents many mechanisms that underlie and cause human behavior including the manner that the operator receives/detects information from an environment, comprehends and registers this information in a memory store, decides on a response, and responds to the information within the context of operational rules and human performance capacities. MIDAS combines these symbolic representations of cognition with graphical equipment prototyping, dynamic simulation, and procedures/tasks to support quantitative predictions of human system effectiveness, and improve the design of crew stations and their associated operating procedures. MIDAS provides an easy to use and cost-effective means to conduct model simulation experiments that explore "what-if" questions about domains of interest. Figure 4.2 illustrates the organization and flow of information among the MIDAS components. For a description of the MIDAS processes, the reader is directed to Gore (2010).

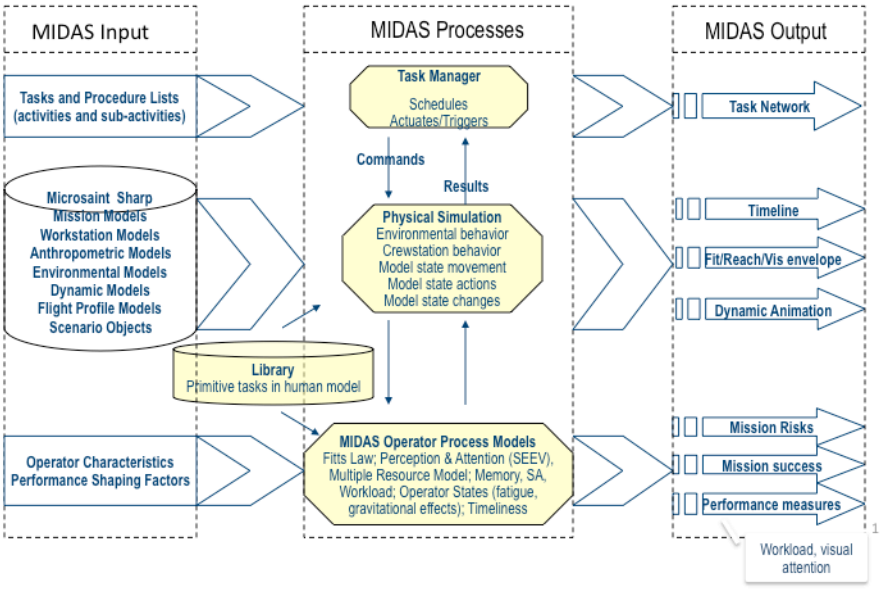

Figure 4.2: Illustration of the flow of information into MIDAS v5.

MIDAS gives users the ability to model the functional and physical aspects of the operator, the system, and the environment, and to bring these models together in an interactive, event-filled simulation for quantitative and visual analysis (Hart et al., 2001). Operator behavior within a MIDAS simulation is driven by a set of user inputs specifying operator goals, procedures for achieving those goals, and declarative knowledge appropriate to a given simulation. These asserted knowledge structures interact with and are moderated by embedded models of perception for extracting information from the modeled world and embedded models of cognition for managing resources, memory, and actions. In this way, MIDAS seeks to capture the perceptual-cognitive cycle of real world operators who work towards their most immediate goals given their perception of the situation and within the limitations of their physical and cognitive capacities. In MIDAS, as in the real world, perceived changes in the world - new information or events - may cause changes in the adjudged context of the situation triggering new goals or altering methods to achieve current goals. Such perceived changes may be precipitated through the behavior of other modeled entities or by user specified time-based, condition-based, or probabilistic events set within the simulation (e.g., a system failure or the receipt of an incoming air traffic control clearance). It may, in fact, be the impact of the operator's own actions which lead to changes in context and goals.

In sum, MIDAS v5 can be used to: develop guidelines (e.g. for NextGen aviation roles and responsibilities and information requirements), design and test procedures, conduct "what-if" scenario evaluations, design and evaluate cockpit designs and information placement, evaluate multi-crew performance designs, predict situation awareness profiles of multiple operators, generate stochastic human performance, predict attention drivers, predict cockpit-related attention overload, generate human operator workload timelines, and determine optimal strategies for human-automation interaction.

Inputs: Data from psychology experiments for the empirically-driven operator models (human perception, cognition, memory, decision, and behavioral models), task network models input into a discrete event simulation tool to represent the tasks and activities required of the human in specific contexts, visualizations of anthropometric characters and of the flight deck and other crewstation

assumptions (e.g., using a computer aided design [CAD] model for a Boeing 747 cockpit), display / environmental information context weights (to populate the attention and situation awareness models).

Outputs:
Overt/observable behavior, time to perform tasks, task accuracy, and task completion / success / failure, human error given task failure, operator workload (visual, auditory, cognitive spatial, cognitive verbal, fine motor, gross motor, and vocal), task/environmental timelines, current ongoing tasks, situation awareness profiles, the task network of the required operator actions, CAD of candidate cockpit designs, operator strategies for human automation interaction and integration.

4.5.2 What Has It Been Applied To?

Primarily to civilian cockpit design, but applications of MIDAS have also been made to space systems (crew vehicle exploration, ISS interface, robotic arm control), military helicopters, and air traffic control. The full list of MIDAS application domains can be found online at http://humansystems.arc.nasa.gov/groups/midas/applications.html while the publications reporting on these and on the empirically driven operator models can be found online at http://humansystems.arc.nasa.gov/groups/midas/publications.html

4.5.3 Where Can You Get It?

A software usage agreement with NASA Ames needs to be obtained to use the software (MIDAS NASA Civil Servant POC: David Foyle – David.C.Foyle@nasa.gov). The MIDAS software operates on the Windows NT platform and requires the freely available .net framework available from the Microsoft website.

4.5.4 How Usable Is It?

MIDAS has proven to be a flexible and useful tool for representing humans operating in a variety of environments. Its greatest strength lies in its ability to represent conceptual designs or candidate procedures in software and then allow a designer or potential user see them operated by a virtual crewmember in the context of a simulation of the target environment. The graphical user interface, timeline output, and anthropometric visualization have made MIDAS accessible to a broader range of potential users. The code has been used extensively over the previous four years. However, from the perspective of a formal usability evaluation determination, given that the MIDAS v5 user guide is currently under development, it is best to consider MIDAS an internal R&D software tool. The software can be downloaded onto a Windows NT platform from a website.

4.5.5 How Extensively Validated Is It?

MIDAS has undergone extensive verification and validation efforts. Each of the component models of human behavior has been validated through empirical human simulation comparison. The component models were then computationalized (represented in algorithmic form) and verified within the MIDAS architecture through extensive verification exercises. The integrated architecture (which contains the host of validated component models) has been validated in (Gore Hooey, Socash et al., 2011).

4.6 Multiple Resource Model

4.6.1 What Is It?

The multiple resource model predicts the success or failure of multi-tasking in the cockpit. The architecture of the multiple resource model is generally built around the multiple resource theory of multi-task performance (Wickens, 1984; 1990; 2005; Wickens, Bagnall, Gosakan, & Walters, 2011, North & Riley, 1989). It is an analytic model that contains two macro elements.

- A **conflict matrix** that predicts the effect of competition for specific (multiple) resources between two time-shared tasks.
- A **demand component** that predicts the interfering effect of the difficulty of the two tasks.

Each is described in turn.

The conflict matrix. The multiple resource model (Wickens, 2008b) assumes that human information processing depends upon some or all of levels along four dichotomous dimensions, as shown in the table. Table 4.2.

Table 4.2. Dimensions of the multiple resource model (modalities now also includes tactile in the latest version)

DIMENSION	**Level 1**	**Level 2**
Stages	Perceptual-cognitive	Action selection
Codes	Verbal/linguistic	Spatial
Modalities	Visual	Auditory
Visual channels	Focal	Ambient.

Note that there is some nesting within these. For example the two visual channels are nested only within the visual modalities, and the different modality channels are nested only within perception/cognition. Note also that the latest version of the multiple resource model now contains Tactile, as a 3rd level of Modalities (Wickens, Hollands et al., 2012)

Any given task can be identified as occupying one or more cells of the matrix. For example talking to ATC will occupy action selection and verbal/linguistic. Thus any pair of tasks may vary in the extent to which they occupy the same cells or different cells. The key to computing resource conflict, is that two tasks will interfere with each other to the extent that they occupy more overlapping cells. The computation mechanism of such prediction is based on the conflict matrix between the tasks. To simplify, if only two resource dimensions were specified (stages and codes) such a conflict matrix would be shown as in Table 4.3

Table 4.3. Representation of the conflict matrix in a simplified four-resource version of the multiple resource model.

Task A / Task B	Perceptual-cognitive spatial	Perceptual cognitive verbal	Action Selection - spatial	Action Selection - verbal
Perceptual – cognitive spatial	.XX	.YY	.ZZ	.AA
Perceptual – cognitive verbal		-	-	-
Action Selection - spatial			-	-
Action Selection - verbal				-

Then the entries within each cell will show how much conflict there will be within a cell demanded by both the task labeled across the rows, and that labeled down the columns. For example if Task A is responding to ATC (response verbal) and Task B is examining a map for route vectors (perceptual - cognitive spatial) the conflict interference between them will be 0.AA. These interference values are fractions ranging from 1 to 1.0. Details on specific numbers can be found in Wickens, 2002a; 2005). Thus any pair of tasks, whose resource demands are ticked off along across the top, and down the sides, will populate a certain number of cells within the matrix. The conflict values for those populated cells are summed to achieve a total resource conflict score.

Total Demand score

Table 4.4 now depicts the original task outline from Table 4.3, within the framework of an outer row and column. These explicitly label the two tasks (here map reading and target searching) and assign a demand vector to each: map reading is heavily perceptual/cognitive and spatial, but here may involve a simple voice response. Visual search (e.g., traffic search out the window) demands only visual perception. The total task demand is simply the sum of the demands of the two tasks, here 2 + [3+1] = 4. Also in this example it can be seen that the conflict score is [.XX + .AA] (the sum of the two populated cells). Hence the total interference is predicted to be the weighted sum of the demand score and the conflict score.

Table 4.4 The simplified four-resource model depicting the task demand vectors of the two competing tasks, map reading and target search.

	Map read	**3**	**0**	**0**	**1**
Target search		Perceptual-cognitive spatial	Perceptual cognitive verbal	Action Selection - spatial	Action Selection - verbal
2	Perceptual – cognitive spatial	**.XX**	.YY	.ZZ.	**.AA**
0	Perceptual – cognitive verbal		-	-	-
0	Action Selection - spatial			-	-
0	Action Selection - verbal				-

Note that in this computational architecture, the two components are treated independently and do not interact. That is, the ".XX" value in the upper left had cell, will be the same with the demand vectors of 2 and 3 (as shown) as it would be if these were, for example, 1 and 1 (two very easy perceptual tasks). Riley et al. (1991) have provided data to show that this simple non-interacting version is fully adequate.

Within the model, the conflict values within the matrix are relatively stable and fixed within the model. However a task analysis is required by the model user to establish which resources are demanded by each task, and the degree of demand within each resource. The latter may be estimated by SMEs, or in some cases, may be available from table lookups, such as the McCracken and Aldrich scale.

The depicted model is analytic, and its computation is straightforward. However two elaborations can be made to provide a discrete event simulation model:

1. As noted in the TLAP architecture (See 4.7 below), if task times are variable, then the overlap between two concurrent tasks will vary according to task time distributions. Then multiple resource conflict and demand should only be computed during the periods when the two tasks overlap in time. During epochs of single task performance, multiple task interference will not (by definition) be occurring.

2. This version of the model assumes a level of interference as computed above. But in reality, if a pilot is confronted with two tasks, whose demand, or resource conflict is so high that it is impossible to process them concurrently, then, by necessity, she must shed one or the other. This task shedding behavior is based on a red-line assumption which is not itself an inherent part of the multiple resource model, and is addressed instead in the architecture of scheduling models such as the scheduling module for task management within MIDAS (Gore, 2013, in preparation; see Section 4.5).

4.6.2 What Has It Been Applied To?

Primarily to flying applications as described in 3.2; although there is one application to driving with in-vehicle technology (Horrey & Wickens, 2004).

4.6.3 Where Can You Get It?

The multiple resource model is not available as a stand-alone software package. However versions of this model have been incorporated into MIDAS (see 4.3 above), and into IMPRINT. IMPRINT is available (upon request) to U.S. Government agencies at the following URL: http://www.arl.army.mil/www/default.cfm?page=445.

4.6.4 How Usable Is It?

Approximations to computing the two components (conflict and demand) are relatively easy to accomplish. Demand values can be assigned by SME's. The demand vectors associated with each task can also be assigned by SME's. Different sets of conflict values within the conflict matrix are available in the published literature (e.g., Wickens, Bagnall et al., 2011).

4.6.5 How Extensively Validated Is It?

Several flight deck validations of the computational model were described above. A single driving validation is described in Horrey and Wickens, (2004). Other less quantitative validations of multiple resource predictions in a flight deck environment have been carried out by Wickens, Sandry, and Vidulich (1983; military cockpit), and Wickens, Goh et al (2003; GA cockpit with NextGen technology).

4.7 Optimal Control Model (OCM) & Flight Control Workload

4.7.1 What Is It?

The Optimal Control Model (OCM; Rikard & Levison, 1981) is actually a prototype of a more general class of manual control models that have been employed (and validated) to predict flight handling qualities (see Wickens, 1986); and the latter commodity is closely related to subjective workload, or more specifically psychomotor load (within a VACP context).

The general architecture of OCM (Kleinman, Baron, & Levison, 1971; Levison, 1989) is to model three fundamental stages of pilot information processing on the flight deck. These are:

1. ***Estimating*** the state of the vector of displayed variables necessary to control an axis of flight (e.g., vertical flight, lateral flight). An optimal Kalman filter is incorporated to do this estimation.

2. ***Predicting*** the future level of these variables to the extent they are subject to lag, and that stable control depends on predicted state, rather than current state.

3. ***Translating*** the estimated predicted state into a vector of control gains, applied to different flight controls. This gain matrix is tuned to *optimize* the balance between the need for precise control (minimizing error) and "smooth flight" (e.g., minimizing large rapid control movements).

All three stages contain optimization components, hence the name "optimal" control.

Importantly, these processes are subject to, and hence modeled with:

- Inherent lags in the human processing system (default, around 1/3 second)
- Noise added to the estimation and control processes.

The most important noise is that added to estimation and observation, (called **equivalent observation noise**) and is assumed to grow linearly as **visual attention is allocated away from the controlled variable in question** (Levison, Elkind, & Ward, 1971), This function provides an important mechanism for multi-task performance and workload prediction.

4.7.2 What Has It Been Applied To?

The model has been primarily applied to flight handling qualities, although one application has been to driving (Levison, 1989).

4.7.3 Where Can You Get It?

The model was at one time available from Bolt, Beranick, and Newman. It is not clear if it still is available.

4.7.4 How Usable Is It?

The OCM is based on the frequency domain language of linear and non-linear feedback control theory. Hence it is designed to work on analog computers (or digital simulation of linear dynamic systems in the frequency domain). As a consequence, its use requires some degree of specialized knowledge in control theory, generally taught within the discipline of aeronautical engineering.

4.7.5 How Extensively Validated Is It?

The model has been validated in a variety of applications to predict flight control performance and workload. The Rickard and Levison (1981) study cited is prototypical of a number of others reported in various proceedings of the Annual Conference of Manual Control.

4.7.6 Extensions of OCM

Other models of manual control within the frequency domain have also been produced and validated to predict flight path tracking performance and occasionally workload. Perhaps the most common of these is the **Crossover model** of McRuer and Jex (1967). We do not review this model here because it is not really a model of human performance, but rather, a model of the whole human-aircraft control system. However we do note that one important feature of this model (and other models of manual control), is the close correlation found between model-predicted lag of an aircraft axis, and measures of handling qualities (assessed by the Cooper-Harper rating scale) which we noted above, is a proxy for the subjective estimate of psychomotor workload. It is assumed that when there is lag along flight control axes, the pilot must “generate lead” (e.g, predict) in order to compensate for the lag and retain flight stability. Prediction is workload-demanding (Wickens, Gempler, & Morphew, 2000) and the greater the need for prediction in flight control, the greater the workload (McRuer & Jex, 1967; Wickens, 1986; figure 39.31).

4.8 The Traffic Organization and Perturbation Analyzer (TOPAZ)

4.8.1 What Is It?

The Traffic Organization and Perturbation Analyzer (TOPAZ) is a multi-agent dynamic risk model designed to evaluate system safety within Air Traffic Management (ATM). The system uses a dynamic extension of Petri nets and Monte Carlo simulations to analyze the safety of air traffic systems through the generation of conditional collision risks. TOPAZ can account for both nominal and non-nominal events and can represent the dynamic interactions between human operators, technical systems and procedures.

TOPAZ has been under development for a number of years. The initial system included a simulation environment based on high-level petri nets. This included support for developing petri net modules for human behavior, the environment and systems within the ATM environment. Additions have included multi-agent situational awareness modeling, support for bias and uncertainty assessment, and modules for accident model specification (Stroeve et al., 2009).

The system supports the development of *systemic accident models* describing the performance of the system as a whole in which accidents become emergent properties of the variability within the system (Stroeve et al., 2009). This quality is important for NextGen in that it supports the evaluation of very low frequency events, such as runway incursions, for which little accident data exists.

The process of developing the model involves representing the dynamic characteristics of related agents, such as human operators and technical systems, within a hierarchical structure. These model structures can represent:

- Key aspects of agents including SA, task scheduling, flight phases, alerting systems
- Agent modes related to human and system performance
- Time characteristics of tasks or systems such as event distributions
- Agent interactions such as task transitions or system availability
- Interactions including the effects of SA on performance and the effect of system availability

(Stroeve et al., 2009)

Numerous Discrete Event Simulation Monte Carlo runs of the model then provide conditional probabilities of incident occurrence based on the dynamic interactions of agents and interactions represented. The TOPAZ system also supports both bias and uncertainty analysis allowing the user to assess the relative contribution of agents and events to the overall system risk.

4.8.2 What Has It Been Applied To?

The primary use of TOPAZ has been the evaluation of collision probability within the ATM environment. As mentioned in the Error (3.1) and SA (3.3) sections, the most common scenario evaluated has been runway incursions resulting from one taxiing aircraft erroneously crossing a runway in which another aircraft is on a take-off roll. The model includes dynamic representations and interactions of the following agents:

- The motions of the take-off and taxiing aircraft through the take-off roll and approach to runway intersection respectively
- The availability and dynamics of the surveillance system including relative position and velocities of each aircraft and the resulting alerts and alarms
- The pilots flying the two aircraft and a range of performance dynamics such as visual monitoring, conflict detection and reaction
- The runway and taxiway controllers and performance dynamics such as communication, conflict detection and reaction.
- The availability and dynamics of the communications systems between the controllers and pilots including delays, frequency selection and the nominal state of these systems.

Figure 4.3 shows the relationships between the agents. The scenario begins with the take-off roll and taxiing of the two aircraft (nodes E10 & E12) and ends with the probability of detection and avoidance or the collision. In one analysis, a huge number of Monte Carlo runs were used to generate the probability of the collision event. This included sensitivity analysis removing agents or varying aspects of agent behavior to assess the effects on the overall risk.

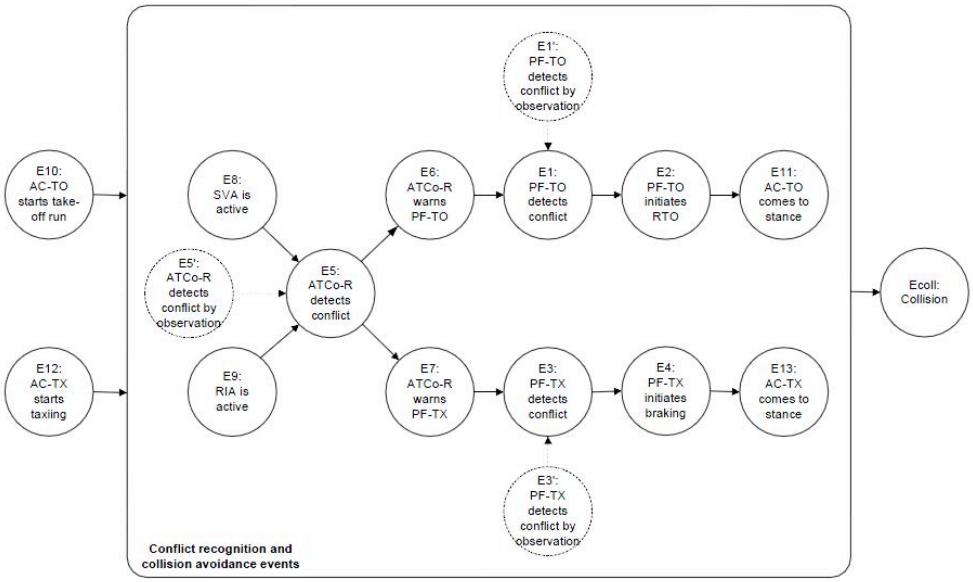

Figure 4.3. Shows the relations events for conflict recognition and collision avoidance actions by the agents in the runway incursion model (from Stroeve et al., 2011).

The results of the bias and uncertainty analysis show that the understanding of pilot performance is the strongest contributor to uncertainty in the risk results. This result would help to direct the analysis team to focus data gathering and modeling efforts on ensuring the accurate understanding and representation of pilot performance within the model. In addition, an analysis of the conditional probabilities of each event given a collision illustrates the individual contributions of each agent to the overall risk. For example, overall collision risk only increases by a factor of 1.06 without an ATC alert system but increases by a factor of 500 if none of the agents (pilots and controllers) are monitoring or in the control loop (Stroeve et al., 2011).

4.8.3 Where Can You Get It?

TOPAZ has been developed by and is available from the National Aerospace Laboratory NLR, Air Transport Safety Institute, Amsterdam, The Netherlands www.nlr.nl.

More information about the model can also be obtained by contacting Henk Blom. (Henk.Blom@nlr-atsi.nl)

4.8.4 How Usable Is It?

TOPAZ is a complex system requiring considerable expertise and effort. The process includes identification of the operation and hazards to be modeled and the development of high-level Petri nets for the various agent models.

4.8.5 How Extensively Validated Is It?

The studies describe a mathematical process of verification rather than validation. Actual data with which to compare model outputs in a true validation are difficult to obtain for such low-probability events.

4.9 Time Line Analysis Procedure (TLAP)

4.9.1 What Is It?

This model has a simple straightforward architecture (Parks & Boucek, 1989). A time line is laid out concerning when different cockpit tasks need to be performed. Each task is assigned a length, which may be fixed (in the analytical version) or have a variance (in a discrete event simulation version). Tasks may overlap in time (as flying the aircraft, while communicating with ATC). Then the total time of a mission is subdivided into time units (e.g., 10 seconds), and within each unit, workload is computed as the ratio of the total of task times within that unit, to the unit length. For example if there are two 5 second tasks within the 10 second unit, the workload is 10/10 = 1.0 (or 100%).

Note that at this simple level, the model does not distinguish whether the two 5 second tasks are performed sequentially or simultaneously. A more refined version of the model adds a penalty for two tasks occupying the pilot at the same time (e.g., simultaneously) and this version has been found to be a more accurate predictor of performance breakdowns (Sarno & Wickens, 1995).

The model can be converted from an analytic model (calculating ratios) to a discrete event simulation model, if the time required by each task has a variance associated with it, and each iteration picks (randomly) a time demand from the distribution for each task. The time line analysis procedure does not incorporate specific assumptions about the pilot response when there is concurrence (e.g., possible shedding or delaying the task, as in a queuing theory model), nor does it consider the nature of the task in question, other than its length. These issues are addressed in the Multiple Resource Architecture.

4.9.2 What has it been Applied To?

As described in section 3.2, it has been applied to many cockpit time-line procedures, with Stone et al. (1987) and Parks & Boucek (1989) offering prototypical examples.

4.9.3 Where Can You Get It?

The original Time Line Analysis (TLA) module was part of a software tool called Crewstation Human Engineering Software System (CHESS), it was used for the Boeing 757/767/747-400 certification. It has not been employed subsequently for certification, and does not currently exist within a commercially available software tool.

4.9.4 How Usable is it?

TLAP is extremely usable, and any user can follow the general guidelines presented in section 3.2.4; or outlined in Parks & Boucek (1989) to create his/her own analysis.

4.9.5 How Extensively Validated is it?

In our review, there are surprisingly few high validity (e.g., commercial flight deck) validations against performance data or task shedding data; given the ease of use. But see Sarno & Wickens (1995) for lower validity validation.

4.10 GOMS

4.10.1 What Is It?

The GOMS (Goals Operators Methods, Selection rules) is an approach to task analysis that is based upon the keystroke level model of human information processing originally crafted by Card, Moran, and Newell (1983). This approach provides fundamental times for basic operations, like looking, item rehearsal, or reaching (as a function of distance and target width). Following a GOMS task analysis, the analyst can assemble tasks so as to create a network of more complex operations, necessary to complete higher-level tasks, like programming an FMS (Polson, Irving, & Irving, 1994; John et al., 2009). The GOMS approach has been used together with the ACT-R modeling architecture to create human performance models that provide predictions for both cognitive errors (ACT-R) and time to complete task sequences (GOMS).

Gil et al., (2012, in press) have enhanced the basic GOMS tool, described in Polson and Javaux (2001), to create E-GOMSL (Enhanced GOMS language), an architecture that is more versatile and appears to be more specifically focused on cockpit tasks. E-GOMSL identifies a set of basic **operators:** [confirm, think-of, look-at, store, recall]. Each operator in turn is characterized by a set of four attributes: [Channel, VACP; Control object (e.g., working memory), syntax, and time (or time distribution)].

Considerable effort in task analysis appears to be required to assemble the operators to approximate the cockpit tasks; but the model will output task times as well as variables such as working memory load, that are important for predicting workload (Gil & Kaber, 2012).

The Model, and the requirements for programming, is well described in Gil and Kaber (2012).

4.10.2 What Has It Been Applied To?

The GOMS approach to task analysis has been applied extensively in analyses of human-computer interaction. It could potentially be used to model any detailed level of interaction in which an agent performs discrete action sequences. There have been some variants of GOMS applied in the aviation environment. These variants include the Cognitive, Performance, Motor (CPM-GOMS). CPM GOMS has focused on F22 operations as well as predicting mouse clicks for controllers

(Remington, Matessa, Freed, & Lee, 2003). A more recent application of the GOMS approach that combines two theory-based tools (CogTool and SANLab) has been used to address human variability in skilled performance as applied to the Boeing 777 Flight Management Computer (FMC) and the Control and Display Unit (CDU) (John, Patton, Gray, & Morrison, 2012). This paper contains no validation data.

4.10.3 Where Can You Get It?

The GOMS method can be applied to any tasks that involve keystroke-level type interactions. Information on this method is available in a variety of sources, including:

Kirwan, B. & Ainsworth, L.K. (1992). *A Guide to Task Analysis*. Washington, DC: Taylor and Francis.

Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human–Computer Interaction*. Hillsdale, NJ: Erlbaum.

http://en.wikipedia.org/wiki/GOMS

4.10.3 How Usable Is It?

GOMS is a potentially time-consuming method for analyzing task sequences, but it can provide useful insights for detailed task sequences. The only usability concerns are related to learning the method itself; there is no particular software that the analyst needs to learn or use.

4.10.4 How Extensively Validated Is It?

GOMS has been in use for 3 decades (as of the date this report was produced) and numerous validations have been conducted using the method (e.g., Gray, John & Atwood, 1993).

5. DISCUSSION and CONCLUSIONS

5.1 Validation Efforts

In evaluating the overall success of the aviation modeling community in generating valid pilot performance models, our findings were somewhat disappointing. In Table 2.4 (see Section 2.2.4), we reported an overall validation rate (across all model aspects) of 36%; and many of the higher values contributing to that overall percentage were aspects not reviewed in our deep dives (e.g., pilot vision models, manual control). Of the six deep dive areas, where we paid close attention to the quality of validation, the percentage was smaller, as shown in Table 5.1, which breaks down validation by whether or not it contained a quantitative aspect.

Table 5.1 Validation statistics for deep dive aspects

Model aspect	All validations	Included Quantitative aspects
Error	9/17	0/17
Workload/multi-task	14/33	11/33
Situation awareness	3/15	1/15
Pilot-automation interaction	8/16	3/16
Roles & responsibilities	3/12	2/12
TOTAL	37/93= 40%	17/93 = 18%

Several reasons can be offered to account for this state of affairs. First, and of relatively minor importance, the statistics reported could be adjusted upward or downward slightly depending on how validation is defined and classified for each model effort. For example some of what we said was ***qualitative*** validation (and hence counted in the left column but not the right) might be argued to be quantitative (e.g., if **some** numerical data were provided from a PITL simulation, but it was challenging to translate these 1-for-1 into the model predictions, e.g., workload was assessed by NASA TLX in an empirical study, yet calculated as VACP scores in the model).

Furthermore, some very good model **verification**, (and we emphasize that a much higher percentage the model architectures reviewed *were* verified, as shown in Table 3.6.1) could be argued by some observers to be a form of ***qualitative validation*** (hence boosting the numbers in the left column of Table 5.1). But at the same time, we might have classified a model as having a qualitative validation, and other observers would challenge this as being a verification, given the fuzziness of the boundaries between these two concepts. Nevertheless we would argue that the precise values of the statistics reported are less important than their approximate level; which indicates that well less than half of the model efforts received validation, and less than half of those received the sort of precise numerical quantitative validation that can both "sell" the model as being accurate (if the correlation is high) and provide precise guidance for model modification, if the correlation is lower. We do observe however in Table 4.1, that all the major architectures have received considerable validation. On this basis, many of the predictions from such architecture made in *model efforts* that have not yet been validated, are still extremely useful because the prior validation of the architecture, provides confidence in the accuracy of the predictions.

To elaborate on the concern about the state of validation, we note that the numerators of several of these statistics in Table 5.1 are heavily populated by studies involving non-transport aircraft and/or non-transport pilots as apparatus or subjects. And even of those studies that used professional pilots as participants, using aircraft simulations representing modern commercial carriers, only a small handful directly examined NextGen issues to test their models predictions.

We offer four non-mutually exclusive reasons for the state of affairs regarding validation.

First, finances for the researcher/modeler are often limited. Most models are developed under contract; and the time and effort to accomplish the full validation at the end of the contract are often underestimated, following the well-known *planning fallacy* (Buehler, Griffin, & Ross, 2002). Funding runs out before complete validation can be accomplished. Furthermore we have noted that in at least two cases, a model was developed in phase 1 of a 3-phase SBIR, and subsequent phases (in which validation was planned) were not funded.

Similarly, along the lines of "limited funding," models provide the opportunity to collect a tremendous variety of data regarding predicted pilot performance. Empirical validation analyses, due to time or resource constraints, sometimes focus on a limited set of the data. Thus sometimes the validation is carried out on only a small sampling of the total data set, such as one set of subjective ratings or one aspect of visual scanning.

Second, validation is cumbersome; particularly in high fidelity PITL simulations. Pilots and simulator time are hard to obtain. Furthermore even when simulations are accomplished, certain kinds of data are hard to quantify, such as "a pattern of errors" (see Section 3.1), or a particular sequence of procedural steps. Also careful experimental design is necessary to create the **different**

experimental conditions necessary to accomplish true validation. That is, is the model sensitive to **changes** in flight procedures or equipment in the same way that pilots are. It is only by comparing model-predicted and pilot-predicted **differences** or **changes** that these critical correlational measures of validation can be obtained.

Third, some validation studies appear to report narrowly focused, promising results rather than a complete set of results. The more comprehensive pilot models (e.g., those created using MIDAS, CASCaS, or ACT-R) can predict many aspects of pilot behavior (e.g., errors, workload, situation awareness), yet several of the validation results provided only data on highly specific issues (e.g., visual workload, errors in selecting a particular page of data). This suggests that the researchers might be reporting only those that indicate good agreement with the model. Alternatively, it might indicate a bias in the literature, where "uninteresting" results (e.g., of a low correlation prediction) are typically not published.

Finally, sometimes numbers may be available from the PITL data, but for reasons that are unclear to the authors of this report, were not converted to appropriate statistics. In this regard we wish to reiterate that the most appropriate statistics are measures of the shared variance accounted for by model and PITL simulation results, where this variance is across-conditions that are meaningful to the FAA (e.g., old vs. new procedures, performance with and without NextGen technology or the 4-way combination of these two). This shared variance is of course represented by the correlation, with N defining the number of conditions compared (see 2.2.3 and Appendix A).

However in this regard we also want to emphasize a point made in Section 2.2, that it is the **raw value** of the correlation (from 0 to 1.0) that is important, not necessarily its statistical significance. This is of course because the latter is influenced by a sample size, and if only 3 or 4 conditions are available for model validation, imposing constraints on statistical significance on a 4-point correlation is an impossibly high bar. Indeed even when only two conditions are available (e.g., conventional vs. NextGen cockpit), and hence a correlation is meaningless, useful quantitative validation can be achieved by comparing the percent *difference* observed in pilot performance (e.g., a 30% reduction in errors) with that predicted by the model. This corresponds to the slope of a regression line, which will be closer to 1.0 as the model is more valid.

5.2. Status of Flight Deck Models

Notwithstanding the limitations of the validation efforts, our team has identified an impressive array of models that either are, or could be tailored to address, issues in NextGen. These models were defined by those "aspects" of pilot performance that they predict, and these aspects could in turn be sorted into three different categories, ordered perhaps from least to most relevant.

First, there were aspects such as spatial disorientation, decision making and manual control that are relevant to all aspects of aviation, and for which there was not time nor resources to pursue in depth. It was decided during our mid-contract meeting with FAA sponsors that these would need to await further study because of our own limited resources.

Second, there were two particular model aspects that we did not initially intend to deep dive, but as we proceeded into the second half of our contract effort we realized were of great relevance to the six aspects that we were pursuing in depth. These were **procedures** and **vision** (particularly **visual attention**). Indeed both of these were partially reviewed in the deep dive effort of Section 3, because

of their high degree of relevance: for example procedural models, such as GOMS are often relevant to activities like programming an FMS, covered in our PAI aspect. Correspondingly, models of visual attention are sometimes a necessary component of situation awareness and automation (PAI) modeling. We believe that these two areas should receive high priority for future evaluation and model development. We also note that vision models in particular were identified early on in our project to have benefitted from a good deal of validation efforts (63% in Table 2.4), and some of this validation was in fact reported in Section 3.

Third of course, are the six model aspects that did receive our deep dive analysis (there were five categories in Section 3, because workload and multi-tasking were combined into one section). After considering these models in some depth, it is our conclusion that, ideally, and at this point in the development of NextGen, they should not be pursued in isolation but rather that architectures should combine aspects in a manner similar to that typified by, for example CASCaS, MIDAS and TOPAZ. Such combination is warranted because the pilots' responsibilities in NextGen are complex and, themselves, integrate these aspects. Several examples abound:

- The pilot using automation (**PAI** aspect) must maintain **situation awareness** of automation state often through **visual attention**
- The pilot's task of programming the **procedures** of complex automation (**PAI**) imposes heavy **workload**
- Different assignment of **roles and responsibilities** can substantially influence **multi-tasking** requirements, as well as impose and/or influence **communications load**.

Each of these three examples link three aspect areas, and for their impact to be predicted, the relevant model should, ideally, incorporate the set of relevant aspects.

Of course with the increasing complexity of multi-aspect models comes the danger of decreased usability for teams and individuals other than the model developer; although it was not our goal in the current effort to "score" models along such a usability scale. Still, we would argue that developers of complex models be sensitive to these issues, and perhaps consider a modular "plug in" concept, whereby a given model aspect, perhaps vital to answer some modeling questions, can have a "default' (and simplified) representation if that aspect is not critical for a particular application. For example a model might have a simplified representation of workload, not requiring channel and resource assignment that could be employed when workload questions are not critical; but could be replaced by a more sophisticated workload model when they are critical

5.3. Final Conclusions

The following list summarizes our main conclusions from this review and analysis:

- There are many modeling architectures available for predicting pilot performance.
- Numerous efforts have been conducted to predict pilot performance.
- There appears to exist a direct relationship between breadth and complexity of modeling efforts (a model that predicts many aspects of pilot performance will probably require extensive modeling expertise).

- Validation efforts are currently insufficient, particularly in model aspects of communications, procedures, and roles & responsibilities, all of which have less than around 25% validation rates.
- Validation efforts should be accompanied by explicit correlation measures, and such measures accompanied by the raw data scatter plots from which they are derived. In this way readers and researchers can inspect the particular conditions (model predictors) that may be "outliers", either under- or over-predicted by the model.
- Verification efforts differ in the extent to which they have been implemented.
- There is a need for clear standards in terms of how models are verified and validated.
- Development of models that address multiple aspects in combination should be encouraged.
- There is a need for more funding for model verification and, especially, validation efforts.

Based on these findings, we recommend the following components for a "highly rated modeling effort": that it include an empirical validation component, that it be scoped appropriately (focused if possible), and that it use an appropriate "tool for the task": one of the broader modeling architectures if it is attempting to model a diverse range of behaviors (e.g., MIDAS, CASCaS, TOPAZ); a cognitive model for addressing cognitive issues (e.g., ACT-R), that it be reasonably well validated. Such an effort should include a clear specification of different NextGen procedures or technologies; and every effort should be made to maintain features of the PITL simulation identical to the conditions of the model simulation. Correlation data should be reported in scatter plot form

We would like to conclude that modeling efforts contribute significantly to improving aviation safety, by providing a viable and useful test bed for evaluating design decisions well before implementation. Modeling efforts provide insights into human cognition and behavior, and offer useful input to system designers

6. REFERENCES

Aldrich, T. B., Szabo, S. M., & Bierbaum, C. R. (1989). The development and application of models to predict operator workload during system design. *Applications of human performance models to system design*, 65-80.

Anders, G. (2001). Pilot's attention allocation during approach and landing: Eye- and head-tracking research in an A330 full flight simulator. Paper presented at the *11th International Symposium on Aviation Psychology*, Columbus, OH.

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Erlbaum.

Banbury, S. & Tremblay, S. (Eds.) (2004). *A Cognitive Approach to Situation Awareness: Theory and Application*. Aldershot: Ashgate.

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.

Boag, C., Neal, A., Loft, S., & Halford, G.S. (2006). An analysis of the relational complexity in an air traffic control conflict detection task. *Ergonomics, 49*, 14, pp 1508-1526.

Boehm-Davis, D. A., Holt, R. W., Chong, R., & Hasberger, T. (2004). Using cognitive modeling to understand crew behavior. *Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA.

Boehm-Davis, D.A., Holt, R.W., Diez, M., & Hansberger, J.T. (2002). Developing and validating cockpit interventions based on cognitive modeling. In W.D. Gray, & C.D. Schunn, (Eds.) *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science*, p. 27.

Boucek Jr, G. P., Sandry-Garza, D. L., & Logan, A. L. Biferno, MA, Corwin, WH and Metalis, S.,(Douglas), 1987. In *Proceedings of the Workshop on the Assessment of Crew Workload Measurement Methods, Techniques, and Procedures: Part Task Simulation Data Summary, AFWAL-TR-87-3103, Sept* (pp. 15-16).

Buehler, R., Griffin, D., & Ross, M. (2002), Inside the planning fallacy: the causes and consequences of optimistic time predictions. In T. Gilovich, D. Griffin, & D Kahneman (Eds), *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, Cambridge, pp. 250-70.

Burdick, M.D., & Shively, R.J. (2000). A full-mission evaluation of a computational model of situational awareness. *Human Factors and Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: HFES

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*. CRC press.

Bzostek, J., Small, R., Bagnall, T., & Walters, B. (2005). *Intelligent Multimodal Signal Adaption System*. Micro Analysis & Design Final Report for NASA-Ames. Contract NNA05AC17C.

Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human–Computer Interaction*. Hillsdale, NJ: Erlbaum.

Carlin, A.S., Alexander, A.L, & Schurr, N. (2010). *Modeling pilot state in next generation aircraft alert systems*. Aptima, Inc.

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: CRC Press.

Corker, K. M., & Pisanich, G. (1998). Cognitive performance for multiple operators in complex dynamic airspace systems: Computational representation and empirical analyses. *Proceedings of the Human Factors and Ergonomics Society, 1*, 341-345.

Deutsch, S., & Pew, R. (2004). Examining new flight deck technology using human performance modeling. *Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA.

Deutsch, S.E., & Pew, R.W. (2008). D-OMAR: An architecture for modeling multitask behaviors. Chapter 8 in D.C. Foyle & B.L. Hooey (Eds.) *Human Performance Modeling in Aviation*. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 183-212.

Donath, D., & Schulte, A. (2009). Behavior model based recognition of critical pilot workload as trigger for cognitive operator assistance. In D. Harris (Ed.): *Engineering Psychology and Cognitive Ergonomics*, HCII 2009, LNAI 5639, pp. 518-528. Berlin / Heidelberg, Germany: Springer-Verlag.

Donnelly, D.M., Noyes, J.M., & Johnson, D.M. (1997). Decision making on the flight deck. *IEE Colloquium on Decision Making and Problem Solving*, pp.3/1-3/4, December 16.

Durso, F.T., Rawson, K., & Girotto, S. (2007). Comprehension and situation awareness. In F. T. Durso, R. Nickerson, S. Dumais, S. Lewandowsky, & T. Perfect, *Handbook of Applied Cognition (2nd),* Chicester: Wiley, pp. 163-193.

Elkind, J.I., Card, S.K., Hochberg, J., & Huey, B.M. (1990). *Human Performance Models for Computer-Aided Engineering*. New York, NY: Academic Press, Inc.

Endsley, M.R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society* 32nd Annual Meeting (Vol. 1, pp. 97-101). Santa Monica, CA: Human Factors Society.

Endsley, M.R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors, 37,* 65-84.

Eng, K., Lewis, R., Tollinger, I, Chu, A., Howes, A. & Vera, A. (2008) Generating automated predictions of behavior strategically adapted to specific performance objectives. *CHI 2008 Proceedings, Automatic Generation and Usability.* Montreal, Can.: Association for Computing Machinery.

Ericsson, K.A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102: 211-245.

Fleetwood, M.D., Lebiere, C., Archer, R., Mui, R., & Gosakan, M. (2006). Putting the brain in the box for human-system interface evaluation. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*. 1165-1169.

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: CRC Press.

Freed, M. (2000). Simulating human agents. *Papers from the 2000 AAAI Fall Symposium*, Michael Freed, Chair. Technical Report FS-00-03. Menlo Park, CA: AAAI Press.

Gil, G., Kaber, D., Kim, S., Kaufmann, K., Veil, T., & Picciano, P. (2009). Modeling pilot cognitive behavior for predicting performance and workload effects of cockpit automation. *Proceedings of the 2009 International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, 124-129.

Gil, G. H., Kaufmann, K., Kim, S. H., & Kaber, D. B. (2010). Effects of modes of cockpit automation on pilot performance and workload in a next generation flight concept of operation.

In *Proceedings of the 3rd International Conference on Applied Human Factors and Ergonomics* [CD-ROM]. Boca Raton, FL: CRC Press

Gil, G.H., & Kaber, D. (2012, in press). An accessible cognitive modeling tool for evaluation of pilot–automation interaction. *International Journal of Aviation Psychology, 22*.

Gonzales-Calleros, J., Vanderdonckt, J., Lüdtke, A. & Osterloh, J.P. (2010). Towards model-based AHMI development. *EICS '10*, June 21-23, Berlin, Germany.

Gore, B.F. (2008). Human Performance: Evaluating the Cognitive Aspects. *Handbook of Digital Human Modeling* (Ch. 32, pp. 1-18).

Gore, B.F. (2010). The use of behavior models for predicting complex operations. *Behavioral Representation in Modeling and Simulation* (BRIMS) 2010. Charleston, South Carolina, Simulation Interoperability Standards Organization (SISO): 1-4.

Gore, B.F. (2013, in prep). *The MIDAS User's Manual*. Moffett Field, CA: NASA Ames Research Center.

Gore, B.F. & Corker, K.M. (2000a). Human performance modeling: Identification of critical variables for national airspace safety. *Human Factors and Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: HFES.

Gore, B.F. & Corker, K.M., (2000b). Value of human performance cognitive predictions: A free flight integration application. *Human Factors and Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: HFES.

Gore, B. F., Hooey, B. L., Haan, N., Bakowski, D. L., & Mahlsted, E. (2011). A methodical approach for developing valid human performance models of flight deck operations. Paper presented at the *Human Computer Interaction International* (HCII) 2011, Orlando, FL.

Gore, B.F., Hooey, B.L., Mahlstedt, E., & Foyle, D.C. (2013). *Evaluating NextGen closely spaced parallel operations concepts with human performance models flight deck guidelines (Part 2 of 2)*. HCSL Technical Report (HCSL-13-02). Moffett Field, CA: NASA Ames Research Center.

Gore, B.F., Hooey, B.L., Mahlsted, E., & Foyle, D.C. (2012). Extending validated human performance models to explore NextGen Concepts. In S. Landry (ed.): *Advances in Human Aspects of Aviation*, pp. 407-416, Boca Raton, FL: CRC Press.

Gore, B. F., Hooey, B. L., Socash, C., Haan, N., Mahlsted, E., Bakowski, D. L., Gacy, A.M., Wickens, C.D., Gosakan, M., & Foyle, D. C. (2011). *Evaluating NextGen closely spaced parallel operations concepts with human performance models*. HCSL Technical Report (HCSL-11-01). Moffett Field, CA: NASA Ames Research Center.

Gore, B.F., Hooey, B.L., Wickens, C.D., & Scott-Nash, S. (2010). *A computational implementation of a human attention guiding mechanism in MIDAS v5.* In V.G. Duffy (Ed.): Digital Human Modeling, HCII 2009, LNCS 5620, pp. 237–246.

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*. Washington, D.C.: National Aeronautics and Space Administration. (NASA/CR-2010-216411).

Gore, B.F., Wickens, C.D., Hooey, B.L., Socash, C., & Gosakan, M. (2012), *MIDAS workload verification: Internal document*. Moffett Field, CA: NASA Ames Research Center.

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: A validation of GOMS for prediction and explanation of real-world task performance. *Human-Computer Interaction, 8*, 3, pp. 237-209.

Hart, S.G., Dahn, D., Atencio, A., & Dalal, K.M. (2001). Evaluation and application of MIDAS v2.0. In the *Proceedings of the Society of Automotive Engineers (SAE) World Aviation Congress*, Sept 2001, Seattle WA (SAE paper 2001-01-2648).

Hooey, B.L., & Foyle, D.C. (2008). Advancing the state of the art of human performance models to improve aviation safety. In D.C Foyle & B.L. Hooey (Eds.), *Human performance modeling in aviation* (pp. 321-349). Boca Raton, FL: CRC Press.

Hooey, B. L., Gore, B. F., Wickens, C. D., Salud, E., Scott-Nash, S., Socash, C., & Foyle, D. C. (2010). Modeling pilot situation awareness. Paper presented at the *Human Modeling of Assisted Technologies Workshop*, Belgirate, Italy.

Hooey, B.L., Gore, B.F., Mahlstedt, E., & Foyle, D.C. (2013*). Evaluating NextGen Closely Spaced Parallel Approach Concepts with Validated Human Performance Models Flight Deck Guideline*s (Part 1 of 2), HCSL Technical Report (HCSL-13-01). Moffett Field, CA: NASA Ames Research Center.

Horrey, W.J. & Wickens, C.D. (2004). Driving and side task performance: The effects of display clutter, separation, and modality. *Human Factors, 46*(4), 611-624.

Hüttig, G., Anders, G., & Tautz, A. (1999). Mode awareness in a modern glass cockpit– attention allocation to mode information. Paper presented at the *10th International Symposium on Aviation Psychology*, Columbus, OH.

John, B. E., Patton, E. W., Gray, W. D., & Morrison, D. F. (2012, September). Tools for Predicting the Duration and Variability of Skilled Performance without Skilled Performers. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 985-989). SAGE Publications.

John, B.E., Blackmon, M.H., Polson, P.G., Fennell, K. & Teo, L. (2009). Rapid theory prototyping: An example of an aviation task. *HFES 53rd Annual Meeting*. *53*(12), 794-798.

Joint Planning & Development Office. (June 2007). *Concept of operations for the next generation air transportation system*, *version 2.0,* Available online at: www.jpdo.gov/library/NextGen_v2.0.pdf, (accessed: 01/21/2009).

Karikawa, D., Takahashi, M., Ishibashi, A., Wakabayashi, T., & Kitamura, M. (2006). Human-machine system simulation for supporting the design and evaluation of reliable aircraft cockpit interface. *SICE-ICASE International Joint Conference*, pp.55-60.

Keller, J., Lebiere, C., & Shay, R. (2004). Cockpit system situational awareness modeling tool. In *Proceedings of the Human Performance, Situation Awareness and Automation Conference* (HPSAA II 2004), Daytona Beach, FL.

Klein, G., Orasanu, J., Calderwood, R., & Zsambok, C.E. (1993) *Decision Making in Action: Models and Methods*. Ablex Publishing Co., Norwood, NJ.

Kleinman, D.L., Baron, S. & Levison, W.H. (1971). A control theoretic approach to manned-vehicle systems analysis. *IEEE Trans on Auto Control*. Vol. AC-16, pp. 824-833, No. 6, December 1971.

Laird, J. E. (2008). Extending the soar cognitive architecture. In *Proceedings of the First Conference on Artificial General Intelligence (AGI-08).*

Laudemann, I., & Palmer, E. (1995). Quantitative analysis of observed workload in the measurement of aircrew performance. *International Journal of Aviation Psychology 5,* 187-197.

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*. Boca Raton, FL: CRC Press.

Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243-1270.

Lee, P.U., Sheridan, T., Poage, J.L., Martin, L., Jobe, K., & Cabrall, C. (2010). *Identification and Characterization of Key Human Performance Issues and Research in the Next Generation Air Transportation System (NextGen)*. NASA/CR-2010-216390.

Leiden, K., Laughery, K.R., Keller, J., French, J., Warwick, W., & Wood, S.D. (2001). *A Review of Human Performance Models for the Prediction of Human Error*. (Technical Report). Boulder, CO: Micro Analysis & Design, Inc.

Levison, W.H. (1989). The optimal control model for manually controlled systems. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R. Sutton, & L. Van Breda (Eds.) *Applications of Human Performance Models to System Design.* (Defense research series, Vol. 2). New York City, NY: Plenum Press. 185-200.

Levison, W., Elkind, J., & Ward, J. (1971). *Studies of multivariable manual control systems: A model for task interference*. NASA Contract report CR 1746. Washington, DC: NASA.

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*.

Lüdtke, A. & Osterloh, J-P. (2010). Modeling memory effects in the operation of advanced flight management systems. Paper presented at the *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.

Lyall, E.A. & Cooper, B., (1992). The impact of trends in complexity in the cockpit on flying skills and aircraft operation. *Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA.

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

McCarley, J., Wickens, C., Sebok, A., Steelman-Allen, K, Bzostek, J., & Koenecke, C. (2009). *Control of Attention: Modeling the Effects of Stimulus Characteristics, Task Demands, and Individual Differences*. NASA NRA: NNX07AV97A.

McNally, B.H. (2005). An approach to human behavior modeling in an air force simulation. *Proceedings of the 2005 Winter Simulation Conference*, pp.5, December.

McRuer, D.T., & Jex, H.R. (1967). A review of quasi-linear pilot models. *IEE Transactions of Human Factors in Electronics*, *HFE-8*(3), 231-249.

Miller, C.A. (1998). *Case Studies Involving W/Index*, Honeywell Technology Center.

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.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review,* 63 (2) 81–97.

Mumaw, R.J., Sarter, N.B., & Wickens, C.D. (2001). Analysis of pilots' monitoring and performance on an automated flight deck. *Proceedings of the 11th biennial meeting of the International Symposium on Aviation Psychology*, Dayton, OH: Wright State University.

Mumaw, R., Boorman, D.J., & Prada, R.L. (2006). Experimental evaluation of a new autoflight interface. Proceedings HCI-Aero 2006, *International Conference on Human Computer Interaction*, September 20-22, 2006, Seattle, Washington.

Muraoka, K., & Tsuda, H. (2006). Flight crew task reconstruction for flight data analysis program. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50*(11): 1194-1198.

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*.

North, R.A., & Riley, V.A. (1989). W/INDEX: A predictive model of operator workload. Applications of human performance models to system design. In G.R.B. McMillan, D.E. Salas. M.H. Strub, R. Sutton, L. van Breda (Eds.) *Applications of Human performance Models to System Design*. Defense Research Series. New York, Plenum Press. 2: 81-90.

Parks, D., & Boucek, G. (1989). Workload prediction, diagnosis and continuing challenges. In G.R. McMillan, D. Beevis ,E. Salas, M.H. Strub, R. Sutton, & L.van Breda. (1989). *Applications of Human performance Models to System Design*. Defense Research Series. New York, Plenum Press. 2.

Pew, R.W., & Mavor, A.S. (1998). *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington, DC: National Academy Press.

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

Polson, P. S., Irving, J., & Irving, S. (1994). *Applications of Formal methods of Human Computer Interaction to Training and Use of the Control And Display Unit*. Tech Report 94-08, University of Colorado.

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.

Raby, M., & Wickens, C.D. (1994). Strategic workload management and decision biases in aviation. *International Journal of Aviation Psychology*. Vol. 4, No. 3, pp. 211-240.

Reason J. (1990). *Human Error*. New York: Cambridge University Press.

Remington, R., Matessa, M., Freed, M., & Lee, S. (2003). Using Apex/CPM-GOMS to develop human-like software agents. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*. Melbourne: ACM Press.

Rickard, W.W., & Levison, W.H. (1981). Further tests of a model-based scheme for predicting pilot opinion ratings for large commercial transports. *Proceedings of the 17th Annual Conference on Manual Control*, pp. 247-256.

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., Young, M.S., Harris, D., Demagalski, J., Marshall, A., Waldmann, T., & Dekker, S. (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.

Salvucci, D.D., & Taatgen, N.A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, *115*, 101-130.

Sarno, K., & Wickens, C. (1995). The role of multiple resources in predicting time-sharing efficiency: An evaluation of three workload models in a multiple task setting. *International Journal of Aviation Psychology*, 5(1), 107-130.

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.

Schoelles. M.J., & Gray, W.D. (2011). Cognitive modeling as a tool for improving runway safety. *The Proceedings of the 16th International Symposium on Aviation Psychology*. Dayton, OH. 541-546.

Schoppek, W., & Boehm-Davis, D.A. (2004). Opportunities and challenges of modeling user behavior in complex real world tasks. MMI-Interaktiv, 7, June, 47-60. ISSN 1439-7854.

Schurr, N. (2011). ALARMS: Alerting and reasoning management system. *Presentation delivered to the 2011 NASA Aviation Safety Technical Meeting*, St. Louis, MO.

Sebok, A., Wickens, C., Leiden, K., Kamienski, J., & Bagnall, T. (2006). *Cockpit-Based Wake Vortex Visualization: Final Report*. Contract No. NNL06AA28P, NASA Langley.

Sebok, A., Wickens, C., Sarter, N., Quesada, S., Socash, C., & Anthony, B. (2012). The automation design advisor tool (ADAT): Development and validation of a model-based tool to support flight deck automation design for nextgen operations. *Human Factors and Ergonomics in Manufacturing and Service Industries, 22*(5), 378-394.

See, J.E., & Vidulich, M.A. (1998). Computer modeling of operator mental workload and situational awareness in simulated air-to-ground combat: An assessment of predictive validity. *The International Journal of Aviation Psychology, 8*(4), 351-375.

Sherry, L., Polson, P., Feary, M., & Palmer, E. (2002) *When Does the MCDU Interface Work Well? Lessons Learned for the Design of New Flightdeck User-Interfaces*. Honeywell Publication C69-5370-0021.

Shively, R. J., Brickner, M., & Silbiger, J. (1997). A computational model of situational awareness instantiated in MIDAS. *Proceedings of the Ninth International Symposium on Aviation Psychology*, Columbus, Ohio.

Stanton, N.A., Salmon, P., Harris, D., Demagalski, J., Marshall, A., Waldmann, T. & Dekker, S. (2003). Predicting pilot error: Assessing the performance of SHERPA. *Proceedings of the HCII Conference*, 587-591.

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

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.

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*.

Stone, G., Culick, R., & Gabriel, R. (1987) Use of task timeline analysis to assess crew workload. In A. Roscoe (Ed) *The practical assessment of pilot workload*. NATO AGARDograph #282.

Sulistyawati, K., Wickens, C.D., & Chui, Y.P. (2011). Prediction in Situation Awareness: confidence bias and underlying cognitive abilities. *The International Journal of Aviation Psychology*, *2*(2), 153-174.

Svensson, E.A.I., & Wilson, G.F. (2002). Psychological and Psychophysiological Models of Pilot Performance for Systems Development and Mission Evaluation. *The International Journal of Aviation Psychology, 12*(1), 95-110.

Tidhar, G., C. Heinze, & Selvestrel, M. (1998). Flying together: modeling air mission teams. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 8*(3), 195-218.

Uijtde Haag, M., Duan, P., Schnell, T., Cover, M., Anderson, N., Snow, M., Etherington, T., Rademaker, R., & Theunissen, E. (2011). *Hazard and integrity monitoring and integrated alerting and notification methods*. Presentation delivered to the 2011 NASA Aviation Safety Technical Meeting in St. Louis, MO.

Walden, R.S., & Rouse, W.B. (1978). A Queueing Model of Pilot Decisionmaking in a Multitask Flight Management Situation. *IEEE Transactions on Systems, Man and Cybernetics* (pp.867-875), December 1978.

Wickens, C.D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and performance* (Vol. 7, pp. 239–257). Hillsdale, NJ: Erlbaum.

Wickens, C.D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of Attention* (pp. 63-101). New York: Academic Press.

Wickens, C.D. (1986). The effects of control dynamics on performance. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of Perception and Performance Vol. II* (pp. 39-1/39-60). New York: Wiley & Sons.

Wickens, C.D. (1990). Resource management and time-sharing. In J.I. Elkind, S.K. Card, J. Hochberg, & B.M. Huey (Eds.), *Human performance models for computer-aided engineering* (pp. 181-202). Orlando, FL: Academic Press.

Wickens, C.D. (2002a). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177.

Wickens, C.D. (2002b). Situation awareness and workload in aviation. *Current Directions in Psychological Science*, *11*(4), 128-133

Wickens, C.D. (2005). Multiple resource time sharing models. In N. Stanton et al. (Eds.), *Handbook of human factors and ergonomics methods* (pp. 40-1/40-7). Boca Raton, FL: CRC Press.

Wickens, C.D. (2008a). Situation awareness. Review of Mica Endsley's articles on situation awareness. *Human Factors, Golden Anniversary Special Issue, 50*, 397-403.

Wickens, C.D. (2008b). Multiple resources and mental workload. *Human Factors Golden Anniversary Special Issue*, *3*, 449–455.

Wickens, C.D., Bagnall, T., Gosakan, M., & Walters, B. (2011). A cognitive model of the control of unmanned aerial vehicles. *The Proceedings of the 16th International Symposium on Aviation Psychology*, Dayton, OH, 535-540.

Wickens, C.D., Gempler, K., & Morphew, M.E. (2000). Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors Journal,* 2(2), 99-126.

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

Wickens, C.D., Harwood, K., Segal, L., Tkalcevic, I., & Sherman, B. (1988). TASKILLAN: A simulation to predict the validity of multiple resource models of aviation workload. *Proceedings of the 32nd Meeting of the Human Factors Society.* Santa Monica, CA: Human Factors Society, 168-172.

Wickens, C.D., Hooey, B.L., Gore, B.F., Sebok, A., & Koenecke, C.S. (2009). *Identifying black swans in nextgen: Predicting human performance in off-nominal conditions*. *Human Factors. 51*(5), 638-651.

Wickens, C.D., Larish, I., & Contoror, A. (1989). Predictive performance models and multiple task performance. *Proceedings of the Human Factors Society 33rd Annual Meeting*, pp 96-100.

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 (A/SA) model of pilot error. In D.C. Foyle & B.L. Hooey (Eds.) *Human Performance Modeling in Aviation*. CRC press.

Wickens, C.D., Sandry, D., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, output, and central processing. *Human Factors*, *25*, 227-248.
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.
Wickens, C.D., Vincow, M.A., Schopper, A.W., & Lincoln, J.E. (1997). Computational models of human performance in the design and layout of controls. *Crew System Ergonomics Information Analysis Center, CSERIAC 97-02.* Dayton, OH: Wright-Patterson Air Force Base.
Yeh, Y., & Wickens, C.D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30* (1): 111-120.
Zacharias, G. L., Miao, A. X., Illgen, C., Yara, J.M, & Siouris, G.M. (1996). SAMPLE: Situation awareness model for pilot in the loop evaluation. *Conference on Situation Awareness in the Tactical Environment*, Naval Air Warfare Center, Patuxent River, MD.

8. APPENDICES

A. MODEL VALIDATION DETAILS (EXTRACT FROM CSERIAC REPORT)
B. REFERENCES INCLUDED IN THE MODEL EVALUATIONS

Appendix A: CSERIAC Report Excerpt

This material is a modified version of Chapter 1 of Wickens et al, 1997.

1. Why Model?

1.1 Early Design Decisions

It is well established that implementing design changes in a system once it has gone into production is very difficult, because of the high cost involved. Thus, human factors deficiencies in a system that are revealed through product evaluation are often discovered too late to be corrected (Elkind, Card, Hochberg, & Huey, 1990). Human factors input to the design process must be provided early in the design cycle, before extensive production setup costs are incurred. But, in the absence of an existing system or prototype configured with a human in the loop to evaluate the design, where should such input originate? A strong case can be made for using computer-based models to "compute" the efficiency of proposed designs from a human factors standpoint (Corker & Smith, 1993). While such models may not be likely to indicate the "best" design, they should be able to raise a red flag by revealing important human factors deficiencies when they are run using a representation of the proposed design as model input. The sorts of human factors deficiencies that directly concern us here are, of course, deficiencies in the layout of information appearing on a display, and, when appropriate, the physical placement of displays and associated controls.

1.2 Conflicting Principles

As those working in the design community will testify, it is rare that a given display arrangement satisfies all human factors layout principles simultaneously. For example, moving one frequently used display close to another frequently used display may mean placing it *farther* away from a display to which it is related. Placing related displays close together may create problems if they are placed *so* close together that the layout becomes cluttered. When two (or more) principles conflict, which one should be followed? Is the best solution a compromise that violates both principles to some extent? A validated model can incorporate answers to such questions and guide the display layout process. A model that takes account of the various principles and provides weights characterizing their relative importance can provide a number that expresses the added benefit of adhering to (or cost of violating) a combination of principles for a given proposed design. Then, the optimal design (or best compromise) can be predicted.

For example, if adhering to principle A (e.g., moving a display closer to another display used in sequence) is twice as important (leads to twice the predicted performance gain) as adhering to principle B (e.g., moving the display closer to a functionally related control), then there is solid justification for choosing a design solution that conforms to A but violates B, even if cost or engineering factors may slightly favor B. In the absence of such models, the designer is left in a state of frustration when a list of sometimes conflicting principles is offered, but no guidance is given on how to resolve these conflicts except by doing further costly experiments. In such instances, it is understandable that the designer will simply choose to satisfy the principles that can be applied most economically. The availability of models should help to address this situation.

1.3 Figure of Merit

A third reason why computational models are important is that they make it possible to assess the degree to which a display layout adheres to human factors principles by providing the designer with

a predicted *figure of merit* for a given display layout. When there is a computational algorithm for measuring the strength of adherence to (or violation of) each principle in isolation, then the overall "goodness" of a display layout can be assessed by properly combining these assessments of adherence to each principle. This combination should, of course, weight the degree of adherence to a given principle by the degree of *importance* of that principle to system performance. In this way, different display layouts can be compared and evaluated before manufacturing begins to determine which one is the "best." Or, alternatively, the impact of a particular design decision (e.g., to reposition a display to a side panel) can be evaluated and its cost or benefit expressed in quantitative terms. Such quantitative comparative data should be useful in trading off human factors constraints against other engineering design or cost constraints. As we have noted, it is particularly valuable to have these data in advance of the actual manufacturing process (i.e., before "metal is bent"), because of the incredibly high cost of making adjustments in design later to accommodate human factors concerns that were revealed only after production had begun (Elkind et al., 1990).

2. Computational Models and Measures: Properties and Criteria

By a computational model, we mean here a tool that "computes," from inputs of parameters that reflect characteristics of a pilot-task-interface description, some numerical index of the quality of performance of the pilot-aircraft system. Such computation is typically (but need not be) done on a computer, in that certain models can predict outputs via a simple algebraic formula (Elkind et al., 1990). A high figure of merit based on a validated model predicts relatively good performance in terms of accuracy and/or speed &/or workload &/or situation awareness. At a minimum, the model should make ordinal predictions (A is better than B is better than C). Preferably, it should be able to make interval- or ratio-scale predictions of how much better A is than B.

Many models are based upon measures of two sorts. There are measures that may define the input to a model, such as the complexity of a particular flight deck operation, or the computed salience of an alert. There are also measures of the model output, performance, workload and/or situation awareness, and these measures, to be useful in model validation, must often be operationalized in particular procedures (e.g, NASA TLX measures of workload; relational complexity as a measure of cognitive complexity. A measure does not become a model, however, until it is incorporated into a quantitative expression that predicts the direction and approximate magnitude of the effects on performance.

There are a number of important criteria by which the value of a model or measure can be judged. The following sections describe the four most important of these criteria: **validation, complexity, practical significance, and a priori specification**. A fifth criterion, usability, is also mentioned, but it is more difficult to apply to the models reviewed in this report.

2.1 Validation

Validating a computational model or measure is very similar in many respects to validating a test. The goal of test validation is to determine if the score earned by an individual on the test correlates with or predicts some *criterion* score measured under other, usually more "operational," circumstances (Anastasi, 1988; Allen & Yen, 1979). For a pilot performance model we ask whether the score (level of performance) predicted for a particular pilot-task-interface combination correlates with the level of performance measured under operational conditions. Do pilots actually perform better in the conditions that are predicted to be better? Do pilots make the same kinds of errors that are predicted?

While we will see that few pilot performance models have been validated in real-world conditions, many have not even been validated against human performance at all. This fact is not meant to be a criticism of their developers; rather, it should serve as a caution to potential users to seek validation of the and also to evaluate the *quality* of whatever validation has been done. In the paragraphs below, we outline some of the factors that should be considered in assessing the quality of empirical validation.

To provide some concrete context for this discussion, suppose a model is developed to predict the optimal layout of cockpit instruments for aircraft flight (Andre & Wickens, 1991). Eight possible layouts are constructed, and the model incorporates parameters that characterize each layout in order to predict a level of performance for each. Pilots then perform a flight task with each layout, a performance score is derived, and a *correlation* is used to evaluate the extent to which the model predictions for each layout match with the obtained data (Figure 1a). The higher the correlation, the better the model and the more successful is the validation. That is, if a model has had its predictions matched against performance, we can say it has been validated. To the extent that the correlation returned by such matching is high, we can say that the validation is **successful.** There are four critical elements involved in creating this correlation and using it to demonstrate model validation: the criterion, the operator sample, the condition sample, and the statistics. We discuss each of these below, to show how their characteristics can influence the validity of the model.

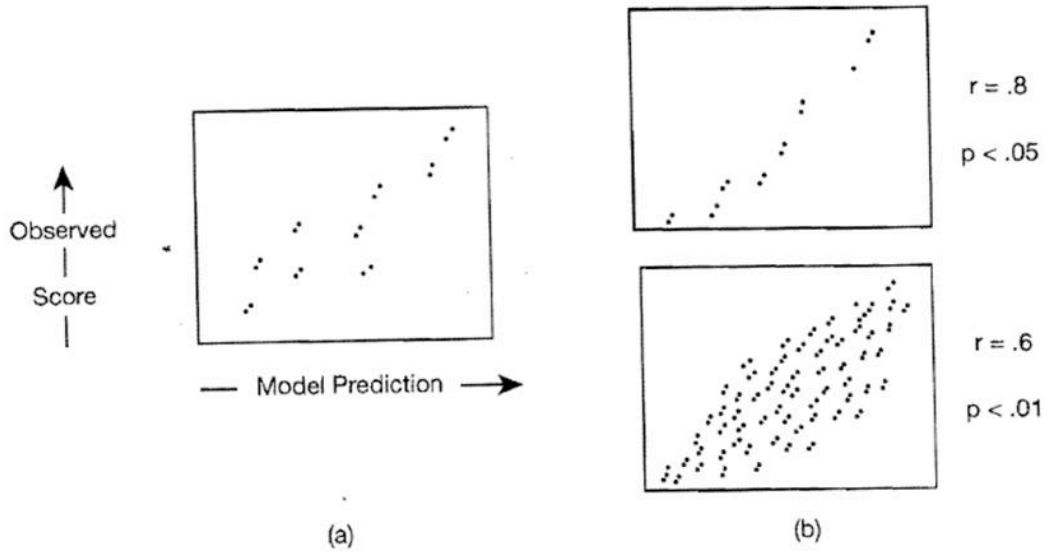

Figure 1. Correlation in model validation.

2.1.1 The Criterion. The criterion variable is the variable that is measured on the y-axis of Figure la. In the Andre & Wickens example, it is a measure of how good a job the display layout does in supporting performance of the task. But what does "good" mean? Ideally, the variable used should be some measure of performance evaluated against objective criteria. In our example, this might include reduced flight path deviation or more rapid and accurate hazard detection. These criterion variables are preferred over operator opinion, simply because in so many circumstances, what users say about an interface does not necessarily agree with how they perform using the interface (Andre & Wickens, 1995).

Ideally, the criterion variable should also be measured in more operational circumstances. We argue that, all other things being equal, measurement in the aircraft is more valid than measurement in the simulator, and measurement in the simulator is more valid than measurement in the basic laboratory setting. But these suggestions do not mean that laboratory studies, or those that have used operator opinion to solicit data on the ordinate of Figure 1a, are "invalid." Our argument is simply that they will be *less robust* than simulator-based and/or real-world measures, all other factors being equal. These "other factors," which are sometimes traded off in favor of more basic laboratory validation, are related to the sample, as discussed next.

The Sample. Each correlation is based upon a set of data points or *cases* (e.g., the points in Figure 1a, which is called the *sample).* The measure of validation is based in part on the identity of those cases. Generally, the greater the degree to which the people whose performance is evaluated are *typical of the real world users* of the system, the greater the generalizability to the actual applied setting. Hence, testing pilots would provide a more meaningful evaluation of a model of cockpit display layout than testing non-pilots.

One reason that validation studies sometimes fail to employ typical system users (experts) is the lack of availability of those experts, which may make it difficult to obtain a *sufficient sample size.* When sample size is small, a correlation can appear to be very high, yet not be statistically reliable, in the sense that it may not be replicated in future studies. Validation studies should seek a large sample, and reports of these studies should always indicate the sample size. Note that if the users are in scarce supply and the real world conditions are expensive to create, it is the need for large samples that sometimes leads model validation efforts away from using typical users and real world conditions.

As noted, the sample is made up of cases. The reported correlation will differ depending on whether those cases, the individual data points in Figure 1a, each represent data that are (a) the average over subjects of performance in each of several test conditions, (b) the individual measure of each subject in each condition, or (c) the average for each subject over several test conditions. Since models are typically developed to predict the effect of different conditions, rather than differences among users, the third option is not appropriate, and is rarely used. The choice between the first two options however is not trivial, however, and which method is used should be clearly reported when correlations are presented.

To illustrate the consequences of this choice, Figure 2 shows validation of a model on four test conditions (A-D) with three subjects (1-3). The raw data points used to compute correlations can represent either the means for the different conditions (averaged over subjects, Figure 2a), the means for the different subjects (averaged over conditions, Figure 2c), or a heterogeneous mixture of the means for each subject and condition (center panel, Figure 2b).

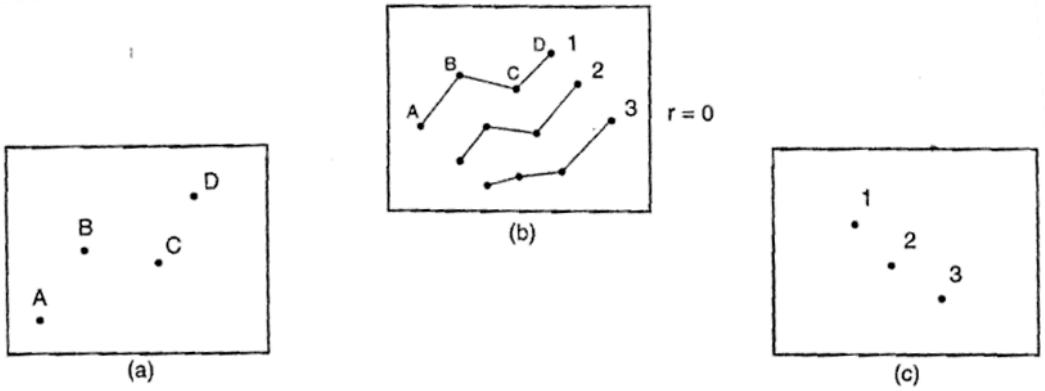

Figure 2. Results of validation study testing three subjects (1-3) under four conditions (A-D). (a) Illustrates a positive correlation between model prediction and obtained data for the four conditions (averaged over subjects). (b) Shows a 0 correlation when variance between subjects is combined with variance between conditions. (c) Shows a negative correlation when data for each subject are averaged over conditions.

As the figure illustrates, each technique can yield a quite different correlation between predicted and obtained scores. Nevertheless, model validation studies are not always explicit about which form of correlation is used. We suggest that the first technique is of greatest value in model validation because it removes variance accounted for by individual differences among subjects, which the model is generally not intended to predict. A plausible alternative is to compute the correlation between model prediction and performance for each subject individually (e.g., correlation underlying the 4 points on each line in Figure 2b) and then report the average correlation across subjects (See Wickens, McCarley, Alexander, Thomas, Ambinder, & Zheng, 2008 for an example). In many respects this approach is optimal because it both shows how well the model fits individual pilot data, as well as the data of the "mean pilot".

2.1.2 The Conditions. Model validation requires the construction of a set of conditions, like the eight display layouts described above, that vary along some parameter(s) incorporated in the model. It is important that the *range of the parameters be appropriate*., neither two wide nor too narrow. Validation efforts may sometimes fall short by creating too little variance between the conditions, relative to the power of the model. For example, if one parameter in the model relates to mean display *separation*, and the validation study used four separations varying in increments of only 0.5 cm, the model might predict little variance in performance. With little variance predicted, a correlation with performance cannot be expected to rise very high, and the validation effort is doomed to failure.

At the other extreme, it is easy to select two or three cases to evaluate that differ by such obvious and excessive magnitudes that variance in performance between them is virtually guaranteed. To take our previous example, we might select display separation differences of 1, 10, and 100 cm. In this case, the differences in visual scanning requirements are large enough to guarantee substantial performance differences (probably lower performance with the wide separation). In this regard, it should also be noted that a single "outlier" point in the appropriate corner of a scatter plot can greatly inflate (Figure 3a) or deflate (Figure 3b) the value of a correlation, particularly if the sample is small, and hence make it appear that the model is doing a much better (or worse) job of predicting performance than it really is. This is illustrated in Figure 3. In 3a, the model appears to have no predictability across the four conditions at the lower left, yet because of the single outlier, the

correlation will be high. In Figure 3, the model does a very good job predicting variance in performance across the four conditions in the lower right, but because of the outlier, the correlation would be very low, and perhaps negative.

The above concerns can be addressed by carefully selecting the parameter values to span the range to be considered by the model (or to be realistically incorporated in actual system design), and, ideally, **by presenting the raw scatter plot** from which the correlations are derived and carefully labeling or identifying any "outlier" conditions that might predict one extreme reading. Correlations should be reported both with and without the outlier, and a clear discussion is needed as to the possible reasons for its removal. That is, what **unique** properties of the condition in question might have caused its displacement from the rest of the data points.

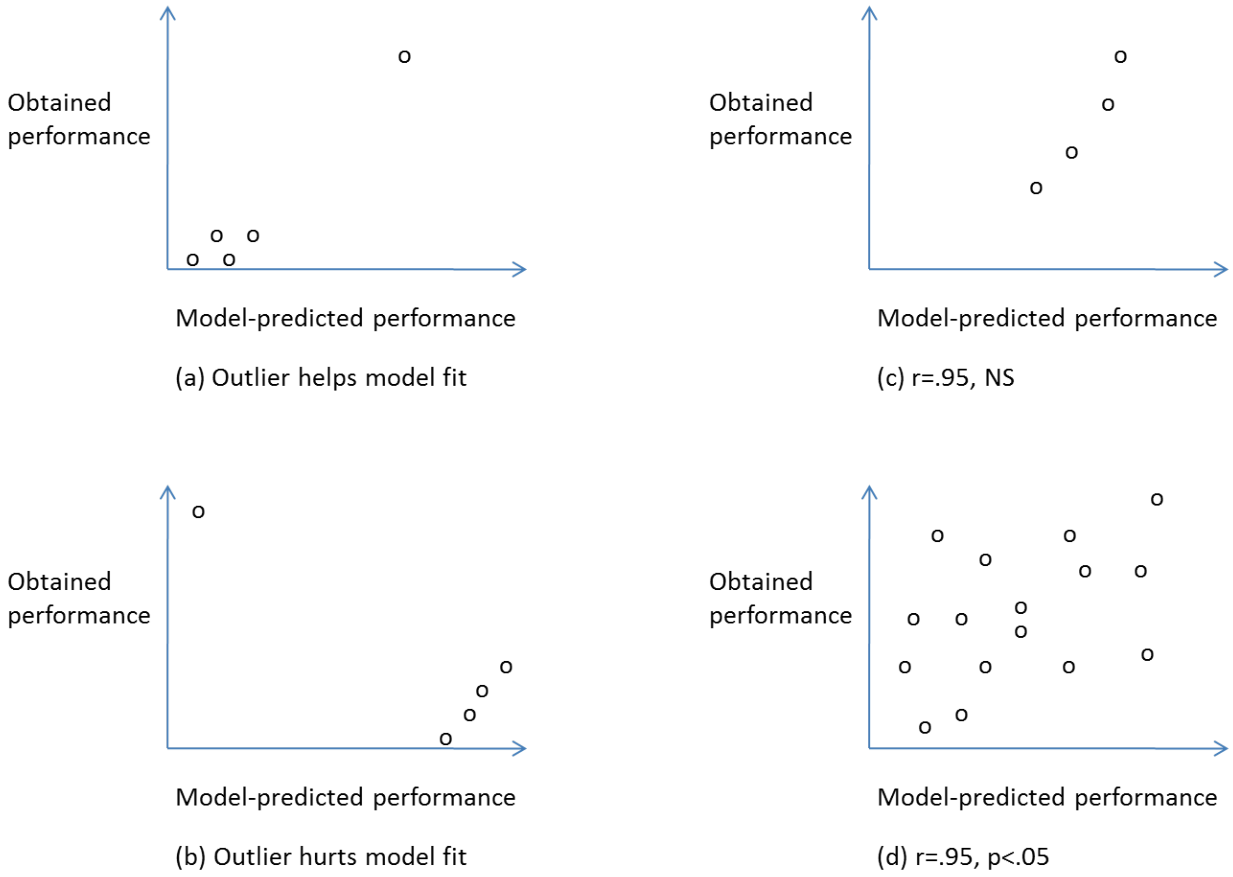

Figure 3. Four examples of model prediction (X axis) vs. obtained performance (Y axis). (See text for a description. The correlation values shown in the figures are not precisely accurate, but are illustrative)

2.1.3 The Statistics. Within the context of a model, the correlation coefficient is a statistic that is often used to characterize the relationship between predicted and obtained scores. The Spearman product moment correlation is the statistic typically used, but rank-order correlations are also sometimes appropriate (particularly when extreme values, as discussed above, may inflate the Spearman value).

Three aspects of the correlation are important to validate a model across conditions: its value (from -1.0 to 0 to +1.0) and its significance (based jointly on its value and the sample size, N, or the number of points in the scatter plot), and N. In classical statistics, most weight is given to significance. But in model validation, equal concern should be given to the value of r. This is illustrated in figure 3c and d. In figure3c the model does a beautiful job of predicting performance, but the high correlation (0.95) may not reach "significance" simply because of the small N. In figure 3d the model is not very good at predicting performance, the correlation is low but it may be significant because of the much larger value of N. Hence authors should always report both the correlation value and its significance and N.

A special case in the statistics of model validation, which is important because of its prominence in many of the models discussed in this report, concerns *linear regression models,* in which the parameters for the model are themselves derived from empirical data. In a typical linear regression approach, the eight display layouts in our experiment might each be characterized by their quantitative level along each of three display variables captured by a model parameter (e.g., average separation, degree of clustering by relatedness, degree of importance). After the dependent variable (e.g., a performance measure) is assessed, linear regression will provide weights dictating the relative importance of each of the three parameters in accounting for variance across the criterion measure (M), such that:

$$M=aA+bB+cC$$

where *A, B,* and C are the levels of each of the three parameters; and *a, b,* and c are the weights of each parameter in the equation.

A well-known characteristic of multiple regression models is that they tend to provide an overly optimistic picture of how well the parameters do in predicting the model dependent variable, because these predictions will capitalize on, and be able to account for, chance or random variation in the criterion measure. This problem can be dealt with in three ways. First, there are objective techniques of "correction for shrinkage" (Tatsuoka, 1971) that reduce the stated estimate of the predicted variance accounted for. Second, and more preferable, is *cross-validation.* In one form of cross-validation, the regression weights are derived on one sample (of conditions, subjects, or both), and then applied to a different sample to determine how well the data of the latter sample are predicted. The latter is then used as the reported validation measure. In another form of cross-validation, the validation experiment is simply repeated. Finally, the third defense against the problem, which is compatible with either of the first two, is to restrict the parameters in a marketed model to those that "make sense" in terms of a conceptual model of human information processing. That is, a parameter might describe a component like working memory load or the breadth of attention that has an independently validated role in human psychology.

All three of these corrective procedures tend to reduce the amount of "significant variance" accounted for by a model and hence, in a way, make the model appear less effective. For this reason, and because of the added complexity of carrying out the first two defenses, many developers of multiple regression models may be reluctant to apply the procedures. But potential model users should be aware of these possible constraints on the validity of regression-based models.

2.1.4 Qualitative Validation. Finally, we note a form of validation in qualitative rather than quantitative form. Here the model predicts some pattern of effects, or distribution of error types, or a

particular error occurring in a particular circumstance. There is no number describing the level of successful prediction, but only data presented that suggest a matching pattern.

3. Complexity

Complexity is a second feature that characterizes the potential value of a model to the user. This feature distinguishes single parameter from multi-parameter models.

Single parameter models may do a precise job of predicting the effect of a given variable (e.g., spatial separation, location in the visual field) on display processing, but they are less than fully satisfactory for use in design. The reason is that they fail to address how the impact of the given parameter will be influenced by other environmental or task characteristics that may vary in real-world conditions. The influence of one model parameter on performance may be greatly affected by (i.e., interact with) another parameter. For example, changes in the spatial separation between two displays will have a very different impact when the displays are related to the same task than when they are not. Hence, it is desirable that models not only recognize the need to address multiple (pertinent) predictive variables but also to include higher-order terms to examine and account for the impact of their potential interactions.

4. Practical Significance of Effects

The value of a model or measure to the user also depends on whether the predictions of the model (or the effects due to the characteristic measured) are on a scale that makes a difference in practice. In operational circumstances, time differences of seconds, or sometimes as short as tenths of a second, are generally of practical importance. In contrast, differences on the order of milliseconds rarely have operational relevance in any sense other than a theoretical one, even though under careful laboratory control small differences may take on a high degree of statistical significance. The magnitude of time difference that is important is, to some extent, context dependent, however. For models predicting the time to perform highly predictable and repetitive operations like key strokes (Card, Moran, & Newell, 1986) differences of tenths or even hundredths of a second can be meaningful.

In general, then, the greater the magnitude of the typical time effects predicted by the model, the greater the value of the model in terms of its *practical significance.* Models that predict errors also score highly on the practical significance criterion.

Note that many validated effects in the millisecond range as demonstrated in the controlled laboratory may indeed turn out to have important design implications. But the *robustness* of these effects in more complex, less controlled environments should be assessed before certifying them as important components of display layout models. Many theoretically important models of visual attention processes have been intentionally excluded from consideration in this report because their practical significance has been deemed low.

5. A Priori Specification

This criterion assesses the degree to which the user of the model or measure can specify model parameters or compute the measure without collecting any new data. For example, some models

require that frequency-of-use estimates for different display components or transition probabilities between displays be obtained before computation of layout design measures can begin (e.g., Freund & Sadosky, 1967; Wierwille, 1981). This feature makes the models more difficult to use. Models and measures that require only information about the physical characteristics of a display are easier to apply than models and measures that require specification of aspects of the task or display content. Models that require formal experiments or observational studies to obtain values for model parameters are of limited usefulness to many designers.

6. Usability

One final criterion that should not be overlooked is the usability of the model or measure for the designer to whom it is recommended. Rouse and Cody (1989) note that designers' use of data sources is heavily dependent on the ease of using those sources. Hence, a well validated model that incorporates many variables, predicts significant performance differences, and requires no data collection to implement may nevertheless be neglected because its features simply make it too difficult for the non-expert to use. Ironically, sometimes the model builder's quest for high validity and explanatory power can make the model so complex (i.e., with too many parameters that must be specified) that it will remain unused. Other features affect usability as well, however. Good human factors at the computer interface, well-written instruction manuals, understanding of the model user's domain, and compatibility with readily available hardware are all examples. While we consider the usability criterion to be critical, it is difficult to apply to most of the models reviewed here, since they are not sufficiently mature for formal computer-based software and user interfaces to have been developed.

7. References

Allen, M.J., & Yen, W.M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks / Cole Publishing Company.

Anastasi, A. (1988). *Psychological Testing* (6th Edition). New York, NY: Macmillan.

Andre, A.D., & Wickens, C.D. (1991). *A computational approach to display layout analysis*. (Technical Report ARL-91-6/NASA-91-2). Savoy, IL: University of Illinois, Institute of Aviation, Aviation Research Laboratory.

Andre, A.D., & Wickens, C.D. (1995). When users want what's not best for them: A review of performance-preference dissociations. *Ergonomics in Design*. October, 10-14.

Card, S.K., Moran, T.P., & Newell, A. (1986). The model human processor: An engineering model of human performance. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.) *Handbook of Perception and Human Performance* (Vol. II, pp 45-1 to 45-35). New York, NY: Wiley.

Corker, K.M., & Smith, B.R. (1993). An Architecture and Model for Cognitive Engineering Simulation Analysis: Application to Advanced Aviation Automation. Paper presented at the *AIAA Computing in Aerospace 9 Conference*, San Diego, CA.

Elkind, J.I., Card, S.K., Hochberg, J., & Huey, B.M. (1990). *Human Performance Models for Computer-Aided Engineering*. New York, NY: Academic Press, Inc.

Freund, L.E., & Sadosky, T.L. (1967). Linear programming applied to optimization of instrument panel and workplace layout. *Human Factors*, *9*, 295-300.

Rouse, W.B., & Cody, W.J. (1989). Designers' criteria for choosing human performance models. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R.Sutton, & L. Van Breda (Eds.), *Applications of Human Performance Models to System Design*. (pp 7-14). New York, NY: Plenum Press.

Tatsuoka, M.M. (1971). *Multivariate Analysis: Techniques for Educational and Psychological Research*. New York, NY: Wiley.

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. Chapter 9 in D.C. Foyle and B.L. Hooey (Eds.) *Human Performance Modeling in Aviation*. Boca Raton, FL: CRC Press.

Wierwille, W.W. (1981). Statistical techniques for instrument panel arrangement. In J. Moraal & K. Kraiss (Eds.), *Manned Systems Design* (pp. 201-281). New York, NY: Plenum.

Appendix B: Sources Included in the Model Evaluations

Al-Zubaidy, S.N. (2008). Proposal for modeling the piloting system. *Second Asia International Conference on Modeling & Simulation*, AICMS, pp.800-805, May 13-15.

Anderson, M.R., Clark, C., & Dungan, G. (1995). Flight test maneuver design using a skill- and rule-based pilot model. *IEEE International Conference on Systems, Man and Cybernetics, Intelligent Systems for the 21st Century*, October 22-25, pp. 2682-2687.

Anderson, M., & Schmidt, D. (1985). Closed-Loop Pilot/Vehicle Analysis of the Approach and Landing Task, pp. 522-526.

Andrews, J.A. (1991). Unalerted air-to-air visual acquisition. Technical report under Air Force contract F19628-90-C-0002. DOT/FAA/PM-87/34, NTIS N9213577. Lexington, MA: MIT Lincoln Laboratory.

Banbury, S., & Tremblay, S. (2004). *A cognitive approach to situation awareness: theory and application*. Ashgate Pub Limited.

Barcheus, F., Ulfvengren, P., & Martensson, L. (2010). Communication enablers for delegation - A relational model for the new ATM system. Paper presented at the *Human Computer Interaction Aero Conference 2010*, Cape Canaveral, FL.

Barnett, B., Stokes, A., Wickens, C.D., Davis, T. Rosenblum, R., & Hyman, F. (1987). A componential analysis of pilot decision-making. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA: HFES.

Baron, S. (1983). An optimal control model analysis op data from a simulated hover task. Proceedings of the *18th Annual Conference on Manual Control*, pp. 195-215.

Baron, S. & Corker, K. (1989) Engineering-based approaches to human performance modeling. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R. Sutton, and L. van Breda (Eds.) *Applications of Human Performance Models to System Design*. (Defense research series, Vol. 2). New York City, NY: Plenum Press. Pp. 203–217

Baron, S., Zacharias, G., Muralidharan, R., Huraldiharan, & Lancraft, R. (1980). Procru: a model for analyzing flight crew procedures in approach to landing. *Proceedings of the 16th Annual Conference on Manual Control*, pp. 495-526.

Bautsch, H., McNeese M. D., & Narayanan, S. (1997). Assessing the value of human performance modeling in exploring pilot-system dynamics. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting.* Santa Monica, CA.

Belyavin, A.J. & Spencer, M.B., (2004). Modeling performance and alertness: The QinetiQ approach. *Aviat Space Environ Med; 75*(3, Suppl.): A93–103.

Belyavin, A., Woodward, A., Nguyen, D., Robel, G., & Woolworth, J. (2005). Development of a novel model of pilot control behavior in balked landings. In *2005 AIAA Modeling and Simulation Technologies Conference and Exhibit* (pp. 1-11).

Benjamin, P. (1970). A Hierarchical Model of a Helicopter Pilot. Human Factors, 12, 361-374.

Besco, R. (1988). Modelling system design components of pilot error. Society of Automotive Engineers Technical Papers. Warrendale, PA. Paper 872517, pp. 53 - 58.

Best, B., Lebiere, C., Schunk, D., Johnson, I., Archer, R. (2005). Validating a Cognitive Model of Approach based on the ACT-R Architecture

Blom, H., Corker, K., Stroeve, S., & Van Der Park, M. (2003). Study on the integration of Air-MIDAS and TOPAZ (NLR-CR-2003). San Jose, CA: NASA/ATAC Corp.

Blom, H.A.P., Corker, K.M., & Stroeve, S.H. (2005). On the Integration of Human Performance and Collision Risk Simulation Models of Runway Operation. In the *6th USA/Europe Air Traffic Management R&D Seminar*, Baltimore, USA, 27-30th June 2005. pp 1-10.

Boehm-Davis, D. A., Holt, R. W., Chong, R., & Hasberger, T. (2004). Using cognitive modeling to understand crew behavior. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting.* Santa Monica, CA.

Boehm-Davis, D.A., Holt, R.W., Diez, M., & Hansberger, J.T. (2002). Developing and validating cockpit interventions based on cognitive modeling. In W.D. Gray & C.D. Schunn (Eds.) *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science*, p. 27

Broussard, J.R., & Stengel, R.F. (1977). Modern control analysis of the pilot-aircraft system. *IEEE Conference on Decision and Control including the 16th Symposium on Adaptive Processes and A Special Symposium on Fuzzy Set Theory and Applications*, pp.235-240, December 1977.

Burdick, M.D., & Shively, R.J. (2000). A full-mission evaluation of a computational model Of situational awareness. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA.

Byrne, M.D., & Kirlik, A.(2004). Integrated Modeling of Cognition and the information environment: A Closed-Loop, ACT-R Approach to Modeling Approach and Lanading With and Without Synthetic Vision System (SVS) Technology.

Byrne, M.D., Kirlik, A., & Fleetwood, M.D. (2008). An ACT-R approach to closing the loop on computational cognitive modeling. Chapter 5 in D.C. Foyle & B.L. Hooey (Eds.) *Human Performance Modeling in Aviation*. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 77 - 104.

Callantine, T.J., (2003). Detecting and Simulating Pilot Errors for Safety Enhancement. SAE Technical Papers.

Carbonell, J. R. (1966). A queueing model of many-instrument visual sampling. *Human Factors in Electronics, IEEE Transactions on*, (4), 157-164

Carlin, A.S., Alexander, A.L., & Schurr, N. (2010). Modeling pilot state in next generation aircraft alert systems. Aptima, Inc.

Chunguang, W., Feng, L., Junwei, H., & Guixian, L. (2008). A Revised Optimal Control Pilot Model for Computer Simulation. The *IEEE International Conference on Bioinformatics and Biomedical Engineering (ICBBE)*, May 16-18

Colle, H.A. & Reid, G.B. (2005). Estimating a mental workload redline in a simulated air-to-ground combat mission. *The International Journal of Aviation Psychology*, *15*(4), 303-319.

Colvin, K., Funk, K., & Braune, R. (2005). Task Prioritization Factors: Two Part-Task Simulator Studies. *The International Journal of Aviation Psychology*, *15*(4), 321-338.

Camacho, R. (1995). Using Machine Learning to extract models of human control skill. *Proceedings of AIT'95*.

Corker, K.M. (2000). Cognitive models and control: human and system dynamics in advanced airspace operations. In N.B. Sarter & R. Amalberti (Eds.) *Cognitive Engineering in the Aviation Domain*. Mahwah, NJ: Lawrence Erlbaum Associates. Pp. 13-42.

Corker, K., H.A.P. Blom, S.H., & Stroeve (2005). Study on the integration of human performance and accident risk assessment models: Air-MIDAS & TOPAZ. *Proceedings of the 2005 International Symposium on Aviation Psychology.* Dayton, OH: Wright State University, 147-152.

Corker, K.M., Muraoka, K., Verma, S., Jadhav, A., & Gore, B.F. (2008). Air MIDAS: A Closed-Loop Model Framework. Chapter 7 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 145-182.

Corker, K.M., & Pisanich, G.M. (1995). Analysis and modeling of flight crew performance in automated air traffic management systems, Oxford, UK, Pergamon.

Corker, K. M., & Pisanich, G.M. (1998). Cognitive performance for multiple operators in complex dynamic airspace systems: Computational representation and empirical analyses. *Proceedings of the Human Factors and Ergonomics Society 1*: 341-345.

Curry, R. E., & Neu, J. E. (1984, September). A model for the effectiveness of aircraft alerting and warning systems. In *Twentieth Annual Conference on Manual Control June 12-14, 1984 Ames Research* (p. 299).

Deutsch, S., & Pew, R. (2002). Modeling human error in a real-world teamwork environment. In W.D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual meeting of the Cognitive Science Society* (pp. 274–279). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Deutsch, S., & Pew, R. (2004). Examining new flight deck technology using human performance modeling. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA.

Deutsch, S., & Pew, R. (2004). Modeling the NASA SVS Part-task Scenarios in D-OMAR. BBN Report No. 8399.

Deutsch, S.E., & Pew, R.W. (2008). D-OMAR: An architecture for modeling multitask behaviors. Chapter 8 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group, pp. 183-212.

Devouassoux, Y., & Pritchett, A. (2001). Application of Kalman filtering to pilot detection of failures. In the *20th Digital Avionics Systems Conference (DASC)*, October 14-18.

Diez, M., Boehm-Davis, D.A., & Holt, R.W. (2002). Model-based predictions of interrupted checklists. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*.

Donnelly, D.M., Noyes, J.M., & Johnson, D.M. (1997). Decision making on the flight deck. *IEE Colloquium on Decision Making and Problem Solving*, pp.3/1-3/4, December 16.

Elkind, J.I., Card, S.K., Hochberg, J., & Huey, B.M. (1990). *Human Performance Models for Computer-Aided Engineering*. New York, NY: Academic Press, Inc.

Emmerson, P. (1997). Worked Example of the Oracle Target Acquisition Model. *Proceedings of the 6th NATO AGARD Meeting*. A6-1 - A6-14.

Eng, K., Lewis, R., Tollinger, I, Chu, A., Howes, A. & Vera, A. (2008) Generating automated predictions of behavior strategically adapted to specific performance objectives. *CHI 2008 Proceedings, Automatic Generation and Usability.* Montreal, Can.: Association for Computing Machinery.

Fotta, M.E., & S. Nicholson (2007). Hemets – Human error modeling for error tolerant systems. *Proceedings of the 14th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, pp 204-209.

Fowler, B. (1981). The aircraft landing test: an information processing approach to pilot selection. *Human Factors, 23*, 129-137.

Frische, F., Osterloh, J.P. & Lüdtke, A. (2010). Simulating Visual Attention Allocation of Pilots in an Advanced Cockpit Environment. *Presented at MODSIM World 2010 Conference Expo*. Pp. 713-729.

Ge., Z., Xu, H., & Liu, L. (2007). A variable strategy pilot modeling and application. *Proceedings of the 2007 IEEE International Conference on Mechatronics and Automation (ICMA)*, August 5 - 8, 2007, Harbin, China.

George, F. L. (1981). Comparison of closed loop model with flight test results. *Proceedings of the 17th Annual Conference on Manual Control*, pp. 296-301.

Gery, K., Doyal, J.,. Brett, B.,.Lebiere, C., Biefield, E., & Martin, E.A. (2003). HPMI: integrating systems engineering and human performance models. *Proceedings of the 12th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, pp 421-426.

Gil, G.H. (2010). An Accessible Cognitive Modeling Tool for Evaluation of Human-Automation Interaction in the Systems Design Process. Unpublished Doctoral Dissertation, North Carolina State University.

Gil, G., D. Kaber, S. Kim, K. Kaufmann, T. Veil, & P. Picciano (2009). Modeling pilot cognitive behavior for predicting performance and workload effects of cockpit automation. Proceedings of the 2009 International Symposium on Aviation Psychology. Dayton, OH: Wright State University, 124-129.

Glenn, F. A. III, & Doane, S.M. (1981). A Human Operator Simulator Model of the NASA Terminal Configured Vehicle (TCV). NASA Contractor Report 3421. NASA Contract NAS1-15983. Langley Research Center, Hampton, VA.

Gonzales-Calleros, J., Vanderdonckt, J., Lüdtke, A., & Osterloh, J.P. (2010). Towards model-based AHMI development. EICS '10. June 21-23, Berlin, Germany.

Gore, B.F. (2008). Human performance: Evaluating the cognitive aspects. *Handbook of Digital Human Modeling* (Ch. 32, pp. 1-18), NJ: Taylor and Francis.

Gore, B. F. (2010). The use of behavior models for predicting complex operations. *Proceedings of the Behavioral Representation in Modeling and Simulation (BRIMS) 2010*. Charleston, South Carolina.

Gore, B. F. (2010). Man-machine integration design and analysis system (MIDAS) v5: Augmentations, motivations, and directions for aeronautics applications. In P. C. Cacciabu, M. Hjalmdahl, A. Lüdtke & C. Riccioli (Eds.), *Human modelling in assisted transportation*. Heidelberg: Springer.

Gore, B.F. & Corker, K.M., (2000a). Human performance modeling: Identification of critical variables for national airspace Safety. In the Human Factors and Ergonomics Society Annual Meeting Proceedings. Santa Monica, CA: HFES.

Gore, B.F. & Corker, K.M., (2000b). Value of human performance cognitive predictions: a free flight integration application. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: HFES.

Gore, B.F. & Corker, K.M. (2002). Increasing aviation safety using human performance modeling tools: An air man–machine design and analysis system application. In M. J. Chinni (Ed.) 2002 *Military, Government and Aerospace Simulation, 34*(3), 183-188. San Diego: Society for Modeling and Simulation International.

Gore, B. F., Hooey, B. L., & Foyle, D. C. (2011, March 21-26). NASA's use of human performance models for NextGen concept development and evaluation. *In* the *20th Annual Conference on Behavioral Representation in Modeling and Simulation 2011 (BRIMS 2011)*, Sundance, UT.

Gore, B. F., Hooey, B. L., Haan, N., Bakowski, D. L., & Mahlsted, E. (2011, July 9 - July 14). A methodical approach for developing valid human performance models of flight deck operations. *In* the *Human Computer Interaction International (HCII) 2011*, Orlando, FL.

Gore, B.F., Hooey, B.L., Mahlstedt, E., & Foyle, D.C. (2013). Evaluating NextGen closely spaced parallel operations concepts with human performance models (Part 2 of 2), HCSL Technical Report (HCSL-13-02). Moffett Field, CA: NASA Ames Research Center.

Gore, B. F., Hooey, B. L., Socash, C., Haan, N., Mahlsted, E., Bakowski, D. L., Gacy, A.M., Wickens, C.D., Gosakan, M., & Foyle, D. C. (2011). Evaluating NextGen closely spaced parallel operations concepts with human performance models. HCSL Technical Report (HCSL-11-01). Moffett Field, CA: NASA Ames Research Center.

Gore, B. F., Hooey, B. L., Wickens, C.D., Socash, C., Gacy, A.M., Brehon, M, Gosakan, M., Foyle, D. C. (2013, in process). *The MIDAS workload model*. HCSL Technical Report. Moffett Field, CA: NASA Ames Research Center.

Goto, N., Chatani, K., & Fuj, S. (1995). H∞-model of the human pilot controlling unstable aircraft. *IEEE International Conference on Systems, Man and Cybernetics, Intelligent Systems for the 21st Century*, pp.2657-2662, October 22-25.

Govindaraj, T., & Mitchell, C.M. (1994). Operator Modeling in Commercial Aviation: Cognitive Models, Intelligent Displays, and Pilot's Assistants. (NASA #NCC 2-675; 90-55). Washington, DC: National Aeronautics and Space Administration.

Griffin, T.G.C., Young, M.S., & Stanton, N.A. (2010). Investigating accident causation through information network modelling. *Ergonomics, 53*(2), 198-210.

Hamilton, D. B., Bierbaum, C.R., & Fulford, L.A. (1991). Task Analysis/Workload (TAWL) (User's Guide) Version 4.0. Fort Rucker, AL.: Anacapa Sciences Inc.

Hayashi, M., Oman, C.M., & Zuschlag, M. (2003). Hidden markov models as a tool to measure pilot attention switching during simulated ils approaches. *Proceedings of the 12th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, pp 502-507.

Heiligers, M.M., Van Holten, T. & Boersema, T. (2003). On a computer-based prediction of pilot scanning workload and control workload. Paper presented at the 12th International Symposium on Aviation Psychology, Dayton, OH.

Heiligers, M.M., Van Holten, T. & Mulder, M. (2009). Predicting pilot task demand load during final approach. *The International Journal of Aviation Psychology, 19* (4), 391-416.

Hess, R. A. (1977). Prediction of pilot opinion ratings using an optimal pilot model. *Human Factors, 19*, 459-475.

Hess, R.A. (1981). Pursuit tracking and higher levels of skill development in the human pilot. *IEEE Transactions on Systems, Man and Cybernetics, 11*(4), pp.262-273, April.

Hess, R. A. (1981). An analytical approach for predicting pilot induced oscillations. *Proceedings of the 17th Annual Conference on Manual Control*, pp. 257-271.

Hess, R. A. (1987). "A Qualitative Model of Human Interaction with Complex Dynamic Systems." *IEEE Transactions on Systems, Man and Cybernetics SMC, 17*(1): 33-51.

Hess, R.A. (2009). Analytical Assessment of Performance, Handling Qualities, and Added Dynamics in Rotorcraft Flight Control. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, pp.262-271, January 2009.

Hill, R.W. Jr. (1999). Modeling perceptual attention in virtual humans. *Proceedings of the 8th Conference on Computer Generated Forces and Behavioral Representation,* Orlando, FL, May.

Hinton, J.L. (1992). The BAE SRC visual performance model - ORACLE an overview (J.S 12138). Filton, Bristol: British Aerospace PLC.

Hoh, R.H., Smith, J.C., & Hinton, D.A. (1987). The Effects of Display and Autopilot Functions on Pilot Workload for Single Pilot Instrument Flight Rule Operations. (NASA Contractor Report 4073). Washington, DC: National Aeronautics and Space Administration.

Holt, R.W., Chong, R., Hansberger, J.T. & Boehm-Davis, D.A. (2002). Modeling crew performance with ACT-R. Technical Report, December 2002. George Mason University, Psychology Department.

Holt, R.H., Chong, R., Schoppek, W., Hansberger, J.T., and Boehm-Davis, D.A. (2002). Modeling crew interaction. Workshop on ACT-R Models of Human-System Interaction.

Hooey, B. L., Gore, B. F., Wickens, C. D., Scott-Nash, S., Socash, C., Salud, E., & Foyle, D. C. (2011). Modeling Pilot Situation Awareness. *Human Modelling in Assisted Transportation*, 207-213.

Hursh, S. R., Balkin, T. J., Miller, J. C., & Eddy, D. R. (2004). The fatigue avoidance scheduling tool: Modeling to minimize the effects of fatigue on cognitive performance. *SAE transactions*, *113*(1), 111-119.

John, B. E., Patton, E. W., Gray, W. D., & Morrison, D. F. (2012, September). Tools for Predicting the Duration and Variability of Skilled Performance without Skilled Performers. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 985-989). SAGE Publications.

John, B.E., Blackmon, M.H., Polson, P.G., Fennell, K., & Teo, L. (2009). Rapid theory prototyping: An example of an aviation task. In the *HFES 53rd Annual Meeting, 53*(12), 794-798.

Jones, R. M., Laird, J. E., Nielsen, P. E., Coulter, K. J., Kenny, P., & Koss, F. V. (1999). Automated intelligent pilots for combat flight simulation. *AI magazine*, *20*(1), 27.

Jonsson, J. E., & Ricks, W. R. (1995). *Cognitive models of pilot categorization and prioritization of flight-deck information* (Vol. 3528). National Aeronautics and Space Administration, Langley Research Center.

Kaber, D.B., Alexander, A.L., Stelzer, E.M., Kim S.H., Kaufmann, K. & Hsiang, S., (2008). Perceived clutter in advanced cockpit displays: measurement and modeling with experienced pilots. *Aviat Space Environ Med, 79*: 1007 – 18.

Kaljouw, W.J., Mulder, M., & van Paassen, M.M. (2004). Multi-loop identification of pilot's use of central and peripheral visual cues. *Proceedings of the AIAA Modelling and Simulation Technologies Conference and Exhibit*. Providence, RI.

Karikawa, D., Takahashi, M., Ishibashi, A., Wakabayashi, T., & Kitamura, M. (2006). Human-machine system simulation for supporting the design and evaluation of reliable aircraft cockpit interface. *SICE-ICASE International Joint Conference*, pp.55-60, October 18-21.

Keller, J., Lebiere, C., & Shay, R. (2004). Cockpit system situational awareness modeling tool. In *Proceedings of the Human Performance, Situation Awareness and Automation Conference (HPSAA II 2004)*, Daytona Beach, FL.

Laudeman, I.V., & Palmer, E.A. (1995). Quantitative measurement of observed workload in the analysis of aircrew performance. *The International Journal of Aviation Psychology, 5(*2), 187-197.

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*. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 105-144.

Laughery, KR (1989). Micro Saint: A tool for modeling human performance in systems. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R. Sutton, & L.V. Breda (eds.) Applications of human performance models to system design (Defense research series, Vol. 2). New York City, NY: Plenum Press.

Levison, W.H. (1989). Alternative treatments of attention-sharing within the optimal control model. *IEEE International Conference on Systems, Man and Cybernetics*, pp.744-749, November 14-17.

Liu, J.Q., & Gao, Z.H. (2010). A test evaluation of a Pilot-Induced-Oscillation prediction criterion. *IEEE 2nd International Conference on Signal Processing Systems (ICSPS),* July 5-7.

Lohrenz, M.C., & Hansman,J., (2004). Investigating Issues of Display Content vs. Clutter During Air-to-Ground Targeting Missions. *In the Proceedings of the Human Factors & Ergonomics Society Annual Meeting* . Santa Monica, CA.

Lüdtke, A. & Osterloh, J-P. (2009). Simulating perceptive processes of pilots to support system design. Human Computer Interaction - INTERACT. *In the Proceedings 12th IFIP TC 13 International Conference Part I*, Uppsala, Sweden, August 24-28, 2009. (pp. 471-484).

Lüdtke, A. & Osterloh, J-P. (2010). Modeling memory effects in the operation of advanced flight management systems. *Paper presented at the 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., & Janssen, J. (2009). Capability test for a digital cognitive flight crew model. *In the Proceedings of the 3rd International Conference on Applied Human Factors and Ergonomics, AHFE 2010* (p. 1-10).

Lüdtke, A., Osterloh, J-P., Mioch, T., Rister, F., & Looije, R. (2010). Cognitive modelling of pilot errors and error recovery in flight management tasks. In the *Proceedings of* 7th IFIP WG 13.5 Working Conference, HESSD 2009, Brussels, Belgium, September 23-25, 2009, Revised Selected Papers, (pp 54-67) .

Lüdtke, A., Weber, L., Osterloh, J. P., & Wortelen, B. (2009). Modeling pilot and driver behavior for human error simulation. *Digital Human Modeling*, 403-412.

Lyall, E. A., & Cooper, B. (1992). The impact of trends in complexity in the cockpit on flying skills and aircraft operation. In the *36 th Human Factors Society Annual Meeting, Atlanta, GA* (pp. 1181-1184).

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

Martin, L., S. Verma, A. Jadhav, V. Raghavan, & S. Lozito (2003). An initial model of data link use in the cockpit. *Proceedings of the 12th International Symposium on Aviation Psychology*. (pp 769-774), Dayton, OH: The Wright State University.

McCoy, M. S. & Levary, R.R. (2000). A rule-based pilot performance model. *International Journal of Systems Science, 31*(6): 713-729.

McMillan, G.R., Beevis, D., Stein, W., Strub, M.H., Salas, E., Sutton, R., & Reynolds, K.C. (1991). A Directory of Human Performance Models (AC/243 (Panel 8)TR/1). Brussels: NATO Headquarters.

McMillan, G.R., Beevis, D., Salas, E., Strub, M.H., Sutton, R., & Breda, L.V. (1989). Applications of human performance models to system design (Defense research series, Vol. 2). New York City, NY: Plenum Press.

McNally, B.H. (2005). An approach to human behavior modeling in an air force simulation. *Proceedings of the 2005 Winter Simulation Conference*, pp.5 pp., December.

Milgram, P., van der Wijngaart R., Veerbeek, H., Fokkerweg, A., & Bleeker, O. (1984). Multi-crew model analytic assessment landing performance and decision making demands. *Proceedings of the 20th Annual Conference on Manual Control, volume 2*, (pp. 374-396).

Miller, C. A., (1998). *Case Studies Involving W/Index,* Honeywell Technology Center.

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.

Mioch, T., Mistrzyk, T., & Rister, F. (2010). Procedure Design and Validation by Cognitive Task Model Simulations. In *Proceedings of the 19th Conference on Behavior Representation in Modeling and Simulation. Charleston, SC, USA* (pp. 232-239).

Mioch, T., Osterloh, J-P, & Javaux, D. (2010). Selecting human error types for cognitive modelling and simulation. *Paper presented at the Human Modelling of Assisted Technologies Workshop*, Begirate, Italy.

Mohlenbrink, C., Lenz, H., Korn, B. (2010). An overview of eye-movement analyses measures for validating a cognitive pilot model. *Paper presented at the Human Computer Interaction Aero Conference 2010*, Cape Canaveral, FL.

Mulgund, S., Rinkus, G., Illgen, C., & Zacharias, G. (1997). Situation awareness modeling and pilot state estimation for tactical cockpit interfaces. HCI International Conference.

Muraoka, K. & Tsuda, H. (2006). Flight Crew Task Reconstruction for Flight Data Analysis Program. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50(*11): 1194-1198.

Muraoka, K., Verma, S., Jadhav, A., Corker, K. M., & Gore, B. F. (2004). *Human Performance Modeling of Synthetic Vision System Use*. Technical Report, San Jose State University, San Jose, CA.

Nelson, W.R., (1988). Functional Models of Complex Human Performance: Application to the Assessment of Pilot Performance. *In the Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Santa Monica, CA: HFES.

Nikolic, M.I., & Sarter, N.B. (2003). Towards a model of error management on highly automated glass cockpit aircraft. *Proceedings of the 12th International Symposium on Aviation Psychology* (pp 882-887). Dayton, OH: Wright State University.

Osterloh, J-P, & Lüdtke, A. (2008). Analyzing the ergonomics of aircraft cockpits using cognitive models. *Proceedings of the 2nd Applied Human Factors and Ergonomics*. 1-10.

Parks, D. & Boucek, G. Workload prediction, diagnosis and continuing challenges. In G.R. McMillan, D. Beevis, E. Salas, M.H. Strub, R. Sutton, & L.V. Breda (1989). *Applications of human performance models to system design (Defense research series, Vol. 2*). New York City, NY: Plenum Press.

Pisanich, G. M., & Corker, K. M. (1995, April). A predictive model of flight crew performance in automated air traffic control and flight management operations. In *Proceedings of the 8th international symposium on aviation psychology* (pp. 335-340).

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

Prasad, S.N. & Schmidt, D.K. (1980). Multi-axis tracking via an optimal control pilot model. *Proceedings of the 16th Annual Conference on Manual Control*, p. 115.

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.

Rao, A. S., Morley, D., Sekvestrel, M., & Murray, G. (1992, November). Representation, selection, and execution of team tactics in air combat modelling. In *Proc. 5th Australian Joint Conference on AI* (pp. 185-190)

Remington, R., Matessa, M., Freed, M., & Lee, S. (2003). Using Apex/CPM-GOMS to develop human-like software agents. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems.* Melbourne: ACM Press.

Rickard, W. W., & Levison, W. H. (1981). Further tests of a model-based scheme for predicting pilot opinion ratings for large commercial transports. *Proceedings of the 17th Annual Conference on Manual Control,* pp. 247-256.

Riley, V., Lyall, E., Cooper, B., & Wiener, E. (1991) Analytic methods for flight-deck automation design and evaluation. Phase 1 report: flight crew workload prediction. FAA Contract DTFA01-91-C-00039. Minneapolis Minn: Honeywell Technical Center.

Ross, L. E., & Mundt, J. C. (1988). Multiattribute modeling analysis of the effects of a low blood alcohol level on pilot performance. *Human Factors, 30*, 293-304.

Rouse, W.B., Hammer, J.M., Mitchell, C.M., Morris, N.M., Lewis, C.M., & Yoon, W.C. (1985). Pilot interaction with automated airborne decision making systems. NASA Grant NAG 2-123. Washington, DC: National Aeronautics and Space Administration.

Rushby, J. (2002). Using model checking to help discover mode confusions and other automation surprises. *Reliability Engineering & System Safety, 75* (2), Feb 2002, pp 167-177.

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., Young, M.S., Harris, D., Demagalski, J., Marshall, A., Waldmann, T., & Dekker, S. (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*,.(pp 567-571).

Sarno, K., & Wickens, C. (1995). The role of multiple resources in predicting time-sharing efficiency: An evaluation of three workload models in a multiple task setting. *International Journal of Aviation Psychology, 5*(1), 107-130.

Schmidt, D. K. (1981). On the use of the ocm's quadratic objective function as a pilot rating metric. *Proceedings of the 17th Annual Conference on Manual Control*, pp. 306-314.

Schoelles. M.J., & Gray, W.D. (2011). Cognitive modeling as a tool for improving runway safety. The *Proceedings of the 16th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, 541-546.

Schoppek, W., & Boehm-Davis, D. A. (2004). Opportunities and challenges of modeling user behavior in complex real world tasks. *MMI interaktiv*, *7*, 47-60.

Schulte, A., & R. Onken (1995). Modeling of pilot's visual behavior for low-level flight. *Proceedings of Synthetic Vision for Vehicle Guidance and Control AeroSense '95*, Orlando, FL (17-21 April 1995).

Schurr, N. (2011). ALARMS: Alerting and reasoning management system. *Presentation delivered to the 2011 NASA Aviation Safety Technical Meetin*g, St. Louis, MO.

Sebok, A., Wickens, C., Sarter, N., Quesada, S., Socash, C., Anthony, B. (2012). The Automation design advisor tool (ADAT): Development and validation of a model-based tool to support flight deck automation design for nextgen operations. *Human Factors and Ergonomics in Manufacturing and Service Industries, 22*(5), 378-394.

See, J.E., & Vidulich, M.A. (1998). Computer modeling of operator mental workload and situational awareness in simulated air-to-ground combat: An assessment of predictive validity. *The International Journal of Aviation Psychology, 8*(4), 351-375.

Shively, R. J., Brickner, M., & Silbiger, J. (1997). A computational model of situational awareness instantiated in MIDAS. .*Proceedings of the Ninth International Symposium on Aviation Psychology*, Dayton, OH: Wright State University.

Siesfeld, A., Curley, R., & Calfee, I. (1984). Communication on the flight deck. *Proceedings of the 20th Annual Conference on Manual Control*, Volume 2, pp.265-275.

Smith, S.C., Govindaraj, T., & Mitchell, C.M., (1990). Operator modeling in civil aviation. *IEEE International Conference on Systems, Man and Cybernetics*, pp.512-514, November 4-7

Sorensen, J., & Goka, T. (1984). Predictions of cockpit simulator experimental outcome using system models, pp 269-290.

Stanton, N.A., Salmon, P., Harris, D., Demagalski, J., Marshall, A., Waldmann, T., & Dekker, S. (2003). Predicting pilot error: Assessing the performance of SHERPA. *Proceedings of the HCII Conference*, pp 587-591.

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

Stokes, A.F. & Raby, M., (1989). Stress and cognitive performance in trainee pilots. *In the Proceedings of the Human Factors & Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: Human Factors Society.

Stone, G., Culick, R. & Gabriel, R. (1987) Use of task timeline analysis to assess crew workload. In A. Roscoe (Ed.), *The practical assessment of pilot workload.* NATO AGARDograph #282.

Stroeve, S.H., & Blom, H.A.P. (2005). Human performance modeling for accident risk assessment of active runway crossing operation. *Proceedings of the 2005 International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, 725-730.

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.

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. In the *9th USA/Europe ATM R&D seminar*.

Stütz, P., & Onken, R. (1997). Adaptive Pilot Modeling within Cockpit Crew Assistance. *Advances in human factors/ergonomics*, 733-736.

Svensson, E., Rencrantz, C., Lindoff, J., Berggren, P., & Norlander, A. (2006, October). Dynamic Measures for Performance Assessment in Complex Environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 24, pp. 2585-2589). SAGE Publications.

Svensson, E.A.I., & Wilson, G.F. (2002). Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. *The International Journal of Aviation Psychology*, *12*(1), 95-110.

Swauger, S. (2003). How good pilots make bad decisions: A model for understanding and teaching failure management to pilots. *Proceedings of the 12th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University, pp. 1137-1142.

Thomas, M. J. W. (2004). Predictors of threat and error management: Identification or core nontechnical skills and implications for training systems design." *International Journal of Aviation Psychology, 14*(2): 207-231.

Tidhar, G., Heinze, C., & Selvestrel, M. (1998). Flying together: Modelling air mission teams. *Applied Intelligence*, *8*(3), 195-218.

Tidhar, G., Selvestrel, M., & Heinze, C. (1995, April). Modelling teams and team tactics in whole air mission modelling. In *Proceedings of the Eighth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE'95)* (pp. 373-381).

Uijtde Haag, M., Duan, P., Schnell, T., Cover, M., Anderson, N., Snow, M., Etherington, T., Rademaker, R., & Theunissen, E. (2011). Hazard and Integrity Monitoring and Integrated Alerting and Notification Methods. Presentation delivered to the 2011 NASA Aviation Safety Technical Meeting in St. Louis, MO.

Van Dongen, H.P.A. (2004). Comparison of mathematical model predictions to experimental data of fatigue and performance. *Aviat Space Environ Med*; 75 (Suppl 1): A15-A36

Verfurth, S.C. Govindaraj, T., & Mitchell, C.M. (1991). OFMspert for the 727: an investigation into intent inferencing on the flight deck. *IEEE International Conference on Systems, Man, and Cybernetics, Decision Aiding for Complex Systems*, pp.1311-1316, October 13-16.

Verma, S., Corker, K. (2002). Introduction of context in a human performance model to predict performance for new air traffic management initiatives. *Proceedings of the Advanced Simulation Technologies Conference 2002*, San Diego, CA.

Verma, S.A., Corker, K. & Jadhav, A., (2003). An approach to modeling error in Air-MIDAS using contextual control model. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*. Santa Monica, CA: Human Factors Society.

Walden, R.S., & Rouse, W.B. (1978). A queueing model of pilot decision making in a multitask flight management situation. *IEEE Transactions on Systems, Man and Cybernetics*. pp.867-875, December 1978.

Washizu, K., Tanaka, K., & Osawa, T. (1980). An experimental study of human pilot's scabning behavior. *Proceedings of the 16th Annual Conference on Manual Control*, pp. 138-144.

Wewerinke, P.R., (1980). The effect of visual information on the manual approach and landing. *Proceedings of the 16th Annual Conference on Manual Control*, pp. 58-74.

Wickens, C.D., Harwood, K., Segal, L., Tkalcevic, I., & Sherman, B. (1988). TASKILLAN: A simulation to predict the validity of multiple resource models of aviation workload. *Proceedings of the 32nd Meeting of the Human Factors Society* (pp. 168-172). Santa Monica, CA: Human Factors Society.

Wickens, C.D., Bagnall, T., Gosakan, M., & Walters, B. (2011). A Cognitive Model of the Control of Unmanned Aerial Vehicles. The *Proceedings of the 16th International Symposium on Aviation Psychology*. Dayton, OH: Wright State University. 535-540.

Wickens, C.D. (1989) Models of multi-task situations. In McMillan, G.R., Beevis, D., Salas, E., Strub, M.H., Sutton, R., Breda, L.V. (1989). *Applications of human performance models to system design (Defense research series, Vol. 2).* New York City, NY: Plenum.

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

Wickens, C. D., Hooey, B. L., Gore, B. F., Sebok, A., & Koenicke, C. S. (2009). Identifying black swans in nextgen: predicting human performance in off-nominal conditions. *Human Factors, 51*, 638-651.

Wickens, C.D., Larish, I. & Contoror, A. (1989). Predictive Performance Models and Multiple Task Performance. Proceedings of the Human Factors Society 33rd Annual Meeting, pp 96-100.

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. Chapter 9 in D.C. Foyle & B.L. Hooey (Eds.) Human Performance Modeling in Aviation. Boca Raton, FL: CRC Press, Taylor & Francis Group. Pp. 213-242.

Wickens, C. D., Sandry, D. L., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *25*(2), 227-248.

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.

Xiaoru, W., Damin, Z., & Hengyang, W. (2009). Pilot attention allocation model in complicated human-machine interface. *International Conference on Biomedical Engineering and Informatics* (BMEI), pp.1-5, October 17-19.

Xiaoru, W., Hengyang, W., & Damin, Z. (2010). Study on pilot attention allocation model based on fuzzy theory. Sixth International Conference on Natural Computation (ICNC), August 10-12, pp. 2035-2039.

Zaal, P. M. T., Pool, D. M., Chu, Q. P., Van Paassen, M. M., Mulder, M., & Mulder, J. A. Delft University of Technology, 2600 GB Delft, The Netherlands.

Zacharias, G. L., Miao, A. X., Illgen, C., Yara, J. M., & Siouris, G. M. (1996). SAMPLE: Situation awareness model for pilot in-the-loop evaluation. *Final Report R*, *95192*.

Zacharias, G., Warren, R., & Riccio, G. (1986). Modeling the pilot's use of flight simulator visual cues in a terrain-following task. *In the 22nd Annual Conference on Manual Control*, pp 81-82, Belton Inn, Dayton, Ohio, July 15th-16th, 1986.

Zuschlag, M., (2004). Quantification of visual cluttering using a computational model of human perception: An application for head-up displays. In *Proceedings of the Human Performance, Situation Awareness and Automation Conference (HPSAA II 2004),* Daytona Beach, FL.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED *(From – To)*
30-04-2013	Technical Memorandum	

4. TITLE AND SUBTITLE

Modeling and Evaluating Pilot Performance in NextGen: Review of and Recommendations Regarding Pilot Modeling Efforts, Architectures, and Validation Studies

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Christopher Wickens, Angelia Sebok, John Keller, Steve Peters, Ronald Small, Shaun Hutchins, Liana Algarin, Brian F. Gore, Becky L. Hooey, David C. Foyle

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

DTFAWA-10-X-80005

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES(ES)

NASA Ames Research Center
Moffett Field, California 94035-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

TH-094

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING REPORT NUMBER

NASA/TM–2013-216504

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified—Unlimited
Subject Category: 1 Availability: NASA CASI (301) 621-0390 Distribution: Nonstandard

13. SUPPLEMENTARY NOTES

Point of Contact: David Foyle, NASA Ames Research Center, 262-2, Moffett Field, CA; (650) 604-3053

14. ABSTRACT

NextGen operations are associated with a variety of changes to the national airspace system (NAS) including changes to the allocation of roles and responsibilities among operators and automation, the use of new technologies and automation, additional information presented on the flight deck, and the entire concept of operations (ConOps). In the transition to NextGen airspace, aviation and air operations designers need to consider the implications of design or system changes on human performance and the potential for error. To ensure continued safety of the NAS, it will be necessary for researchers to evaluate design concepts and potential NextGen scenarios well before implementation. One approach for such evaluations is through human performance modeling. Human performance models (HPMs) offer advantages over empirical, human-in-the-loop testing in that they allow detailed analyses of systems that have not yet been built; offer flexibility for extensive data collection; and they don't require experimental participants. HPMs differ in their ability to predict performance and safety with Nexten procedures, equipment and ConOps. Our research objectives were to support the FAA in identifying HPMs appropriate for predicting pilot performance in NextGen operations, provide guidance on how to evaluate the quality of different models, and to identify gaps inpilot performance modeling research that could guide future research. This research is intended to help the FAA evaluate pilot modeling efforts and select the appropriate tools for future modeling efforts to predict pilot performance in NextGen operations.

15. SUBJECT TERMS Pilot human performance models; NextGen; Human performance

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Help Desk at email: help@sti.nasa.gov
U	U	U	UU	210	19b. TELEPHONE NUMBER (Include area code) STI Help Desk at: (301) 621-0390