A Methodical Approach for Developing Valid Human Performance Models of Flight Deck Operations

Brian F. Gore[1], Becky L. Hooey[1], Nancy Haan[2], Deborah L. Bakowski[1],
and Eric Mahlstedt[1]

[1] San Jose State University at NASA Ames Research Center
MS 262-4, P.O. Box 1, Moffett Field, CA 94035-0001
{Brian.F.Gore,Becky.L.Hooey,Debi.Bakowski,
Eric.Mahlstedt}@nasa.gov
[2] Dell Services Federal Government
MS 262-4, P.O. Box 1, Moffett Field, CA 94035-0001
Nancy.Johnson@NASA.gov

Abstract. Validation is critically important when human performance models are used to predict the effect of future system designs on human performance. A model of flight deck operations was validated using a rigorous, iterative, model validation process. The process included the validation of model inputs (task trace and model input parameters), process models (workload, perception, and visual attention) and model outputs of human performance measures (including workload and visual attention). This model will be used to evaluate proposed changes to flight deck technologies and pilot procedures in the NextGen Closely Spaced Parallel Operations concept.

1 Introduction

The National Airspace System (NAS) in the United States is currently being redesigned because it is anticipated that the current air traffic control (ATC) system will not be able to manage the predicted two to three times growth in air traffic in the NAS [1]. The goal of the Next Generation Air Transportation System (NextGen) is to increase the capacity, safety, efficiency, and security of air transportation operations [1]. However, in doing so, it is expected that the data available to pilots on the flight deck (e.g., weather, wake, traffic trajectory projections, etc.) will be increased substantially in order to support more precise and closely coordinated operations. If not designed with consideration of the human operators' capabilities, these NextGen concepts could leave pilots, and thus the entire aviation system, vulnerable to error.

Human Performance Models (HPMs) have been shown to play a role in all phases of the concept development, refinement, and deployment process of next generation systems [2,3]. HPMs can be used to develop and evaluate new technologies, operational procedures, and the allocation of roles and responsibilities among human operators and automation. HPMs hold the most promise when they are used early in the system design, or system redesign, process and when used iteratively with human in the loop (HITL) simulation output [3,4]. However, before HPMs can be successfully implemented to evaluate how NextGen concepts will impact pilot

M. Kurosu (Ed.): Human Centered Design, HCII 2011, LNCS 6776, pp. 379–388, 2011.
© Springer-Verlag Berlin Heidelberg 2011

performance, baseline models of current-day pilot performance must first be developed and validated.

The objective of this research effort was to develop and validate a baseline HPM of current-day pilot performance. This model will then be used to evaluate proposed changes to flight deck technologies and pilot roles and responsibilities in NextGen Closely Spaced Parallel Operations (CSPO) concepts. Model outputs, including pilot workload and visual attention will be used to draw conclusions regarding the requirements necessary to support NextGen concepts and predict human performance effects, identify safety vulnerabilities, and recommend mitigations.

1.1 Man-machine Integration Design and Analysis System (MIDAS)

The Man-machine Integration Design and Analysis System (MIDAS) is a dynamic, integrated human performance modeling environment that facilitates the design, visualization, and computational evaluation of complex man-machine systems [5]. MIDAS symbolically represents many mechanisms that underlie and cause human behavior. Figure 1 illustrates the model's organization and flow of information among the model's components.

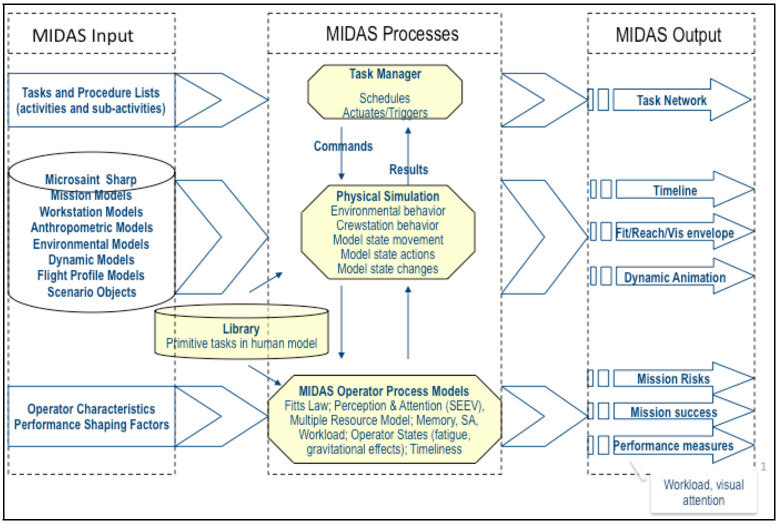


Fig. 1. MIDAS architecture (adapted from [2])

MIDAS inputs (Fig. 1, left column) include the operators' task and procedures, the operational environment (e.g., flight profiles, scenario objects and events, cockpit layout etc), and operator characteristics (e.g., operator expertise, and fatigue).

The *MIDAS processes* (Fig. 1, middle column) are comprised of a task manager model that schedules tasks, definitions of the state of models within the physical simulation, a library of "basic" human primitive models that represent behaviors required for all activities, and cognitive models such as operator perception, visual attention, and workload. These basic process models have been extensively validated.

For instance, MIDAS' attention-guiding model operates according to the SEEV model [6], an extensively validated model that estimates the probability of attending, P(A), to an area of interest in visual space, as a linear weighted combination of the four components - salience, effort, expectancy, and value. The SEEV model has been integrated into MIDAS [7] and drives the operators' visual attention.

Visual perception in MIDAS depends on the amount of time the observer dwells on an object and the perceptibility of the observed object. The perception model computes the perceptibility of each object that falls into the operator's field of view based on properties of the observed object, the visual angle of the object and environmental factors. In MIDAS, perception is a three-stage, time-based model (undetected, detected, comprehended) for objects inside the workstation (e.g., an aircraft cockpit) and a four-stage, time-based perception model (undetected, detected, recognized, identified) for objects outside the workstation (e.g., taxiway signs on an airport surface) [8]. Information then passes into a three-stage memory store [9] that degrades according to empirically-driven memory decay rates [10].

The cognitive models interact with a series of validated anthropometric models that call a number of validated motor movement models [11]. For a description of the MIDAS processes and empirical models, the reader is directed to [5, 12].

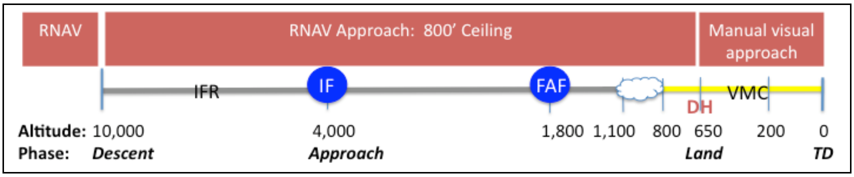
The *MIDAS output model* (Fig. 1, right column) generates a runtime display of the task network, timeline, fit, reach, and visibility envelopes, a dynamic animation of the operator carrying out his/her tasks within the environment, and mission performance measures such as workload and visual fixations.

2 Modeling Flight Deck Operations

2.1 Developing a Model of Current Day Approach and Land Operations

The objective was to develop a high-fidelity model of two-pilot (pilot flying, PF, and pilot-not-flying, PNF) commercial transport operations, with ATC tasks and procedures modeled at a lower level of fidelity, but at a level sufficient to represent the interactions between pilots and ATC. The model was based on a scenario in which pilots flew an area navigation (RNAV) approach into Dallas Fort Worth (DFW) with current-day Boeing-777 equipage (see Figure 2). The scenario began with the aircraft at an altitude of 10,000' and 30nm from the runway threshold. The cloud ceiling was 800', with a decision height of 650' at which point the modeled pilots disconnected the autopilot and manually hand-flew the aircraft to touchdown.

The RNAV model was based on cognitive task analyses of flight tasks [13,14] and cognitive walkthroughs with a commercial pilot and ATC. This process generated a comprehensive set of tasks in each of the following major phases of flight: Descent, Approach, and Land. During descent (10,000' to 4,000'), the PF controls the aircraft autopilot using the MCP and the PNF is primarily responsible for radio communications, checklists, and crosschecking. During approach (4,000' to 650'), the crew configures for the aircraft for landing by progressively lowering flaps and then the landing gear. At the Final Approach Fix (FAF), the PNF radios Tower Control, to obtain landing clearance. In the Land Phase (650' to touchdown (TD)) the crew prepares to land the aircraft. After obtaining a visual identification of the runway, the PF disconnects the autopilot and flies the aircraft to touchdown on the runway.



Notes: DH = Decision Height; FAF = Final Approach Fix; IF = Initial Fix; IFR = Instrument Flight Rules; RNAV = Area Navigation; TD = touchdown; VMC = Visual Meteorological Conditions.

Fig. 2. Baseline RNAV model of approach and land

The task model is composed of major procedures that are then broken down into a set of task primitives at a fine-grained level of fidelity. For example, the task of pressing a button on the MCP is translated into the following sequence of behavioral primitives: *reach, push and release, return arm*. These are then translated into MIDAS' Micro Saint Sharp task network structure. The model was composed of over 970 tasks including environment parameters and flight crew or ATC tasks.

Verifying the Model. To verify that the model was implemented error-free, a new task analysis was reverse-engineered from the model output of task begin and end times and this was compared to the original task analysis. The reverse engineering process culminated in a list of pilot tasks and associated task times and sequence. This was evaluated by an independent pilot, not involved in the initial model development process, for accuracy and completeness.

2.2 Validating the Model of Flight Deck Operations

Model validity can be considered from many different perspectives [4,5] including evaluation of model inputs and model outputs. Model inputs refer to the task trace and model input parameters such as task times and task loads, whereas model outputs refer to operator performance measures, such as visual fixations, workload, or situation awareness. HPMs of complex operations should be evaluated using multiple measures that address varying levels of fidelity. Relying only on output validation, (also referred to as results validation) as the sole measure as is frequently the case is insufficient because there is no guarantee that the model represents the operators' tasks or cognitive processes accurately [2,3,15]. Indeed, it is possible that one could make parameter manipulations until the model output fits the data, while misrepresenting the sequence or order of tasks, the workload associated with carrying out the individual tasks, or the way the operator processes the information from the environment [3]. In this case, the model does not validly represent the pilot's tasks and may lead to invalid conclusions when the model is extended to new scenarios, tasks, or environments.

The MIDAS approach for model validation is presented in Figure 3. This methodical approach is multi-dimensional using multiple variables at varying levels of resolution. The validation process involves three validation components: validation of the inputs, validation of the process models, and validation of the model outputs. The input and output components are scenario-specific while the architecture process models are general and not specific to the domain application model (i.e. perception is perception no matter if the task is driving or flying). Note that validation of the process models themselves, including workload management, perception, and visual attention are important aspects of the validation progression. These process models within MIDAS have been previously validated (as discussed in Sect 1.1) and are held constant across domain applications, so will not be discussed further here.

The application model of flight deck operations was validated by a thorough evaluation of ***model inputs***, including the task trace and model input parameters required for the workload and visual attention models and ***model outputs*** including workload and visual fixation (percent dwell time; PDT). The process was iterative, in that the model was refined based on the input validation process and the output were validated by comparing the refined model to empirical HITL data.

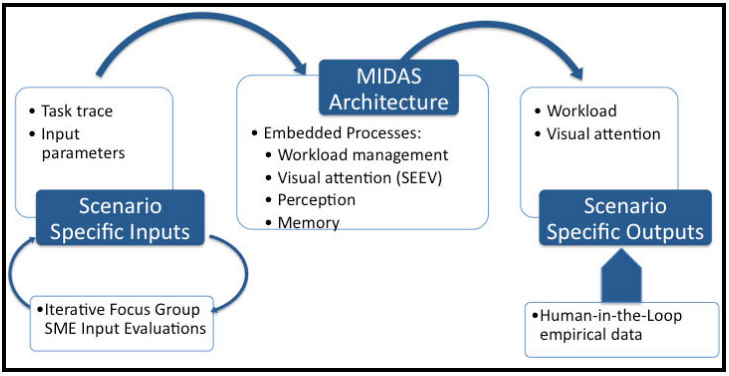


Fig. 3. Comprehensive model validation approach

2.3 Input Validation

Our validation of the model inputs included two aspects. First, a formal validation of the task trace was conducted to determine the extent to which the modeled tasks represent the pilots' actual tasks. Second, a formal analysis was conducted to determine the validity of the model input parameters of workload assigned to the basic task primitives. MIDAS uses behavioral primitives that contain workload estimates based on the Task Analysis and WorkLoad (TAWL) index [16]. These values are based on inputs from military rotorcraft pilots, and have not previously been validated by commercial pilots, for the task of conducting approach and landing tasks in fixed-wing aircraft.

Method. Two full-day focus group sessions were conducted to evaluate the validity of the model inputs. Each session was composed of four pilots. The eight pilots (six Captains and 2 First Officers) were current commercial pilots of glass-cockpit aircraft (M=1,317 flight hours), with RNAV-approach experience. Using a scenario-based format, the focus group pilots conducted a cognitive walkthrough of a typical approach-and-land scenario, starting at 10,000' and continuing to touchdown. Pilots were asked to consider the tasks required of the PF and the PNF, and the nature of communications between ATC and the pilots.

The PF, PNF and ATC tasks as modeled in MIDAS were presented on a worksheet and each pilot was asked to independently review the task and identify any tasks that were assigned to the incorrect operator, occurred in the incorrect sequence, or at an incorrect altitude or navigation marker, and identify tasks that were missing. Upon completion of the worksheet, the pilots discussed their evaluations in a semi-structured round-table format and the source of any discrepancies among pilots was identified and resolved. Differences among tasks were attributed to differences due to aircraft type, airline, pilot technique, and airport/airspace procedures.

Next, the pilots were trained to estimate task workload along five dimensions (Visual, Auditory, Cognitive, Speech and Motor) using the 7-point modified TAWL scale with behavioral anchors [16,17]. The pilots were asked to first identify the workload dimensions that were applicable for the given task, and then estimate the workload for each relevant dimension using the 7-point scale. Two categories of behavioral primitives were evaluated: 1) Basic behavioral primitives existing in MIDAS based on the TAWL that were deemed a valid representation for rotorcraft operations, and, 2) RNAV model-specific behavioral primitives, which had not been previously validated. The basic behavioral primitives served as a baseline upon which to evaluate whether the focus group pilots' workload estimates were comparable to those MIDAS behavioral primitives that were based on the previously validated TAWL scale.

Results

Task Trace Validation. Out of 74 pilot procedure tasks in the MIDAS RNAV application model, the focus group pilots identified 12 tasks that should be removed, reordered, or added. In addition, pilot-ATC phraseology was refined to better reflect actual operations. Incorrectly representing the communication length or the information contained within the communication results in misestimates of workload and task time to reach comprehension. The MIDAS input model was modified to reflect these changes.

Input Parameter Validation. For each task, the mean estimated workload for each workload dimension was compared to the MIDAS input parameter, with the constraint that at least six of the eight focus group pilots determined that the dimension was relevant for the task. One sample t-tests were conducted, which compared the mean focus group rating to the MIDAS value. Significant results indicated that the pilots' estimated workload values were significantly different than the MIDAS values. Thirty-nine tasks were rated on the visual, auditory, cognitive, and motor dimensions (as relevant for the task) resulting in 75 ratings.

Three of the behavioral primitives were established, previously validated, MIDAS behavioral primitives. They were *push-and-release, reach object, and say message*. For all three primitives, the focus-group mean ratings did not differ significantly from the existing MIDAS ratings (*push-and-release* - $t_{visual}(7)=4.2$, $p>.05$, t_{motor} $(7)=12.19$, $p>.05$; *reach object* – $t_{visual}(7)=.306$, $p>.05$, $t_{cognitive}$ $(7)= 1.17$, $p>.05$, t_{motor} $(6)=1.37$, $p>.05$; *say message* - $t_{cognitive}$ $(7)=.877$, $p>.05$). This is evidence that the focus group pilots were trained sufficiently on the TAWL scale to produce answers in accordance with the TAWL, thus providing confidence in the pilots' ratings for the non-TAWL primitives, as discussed next.

The initial model mapped four of the pilots' tasks (*set speed, set flaps, set gear, and tune radio frequency*) to the *push-and-release* task primitive. However, the focus-group ratings of these input parameters revealed that each possessed unique workload properties that differed significantly from *push-and-release* (*set speed* - $t_{visual}(7)=3.5$, $p<.05$; $t_{cognitive}(6)=1.7$, $p<.05$; *set flaps* - $t_{motor}(7)=4.5$, $p<.05$; *set gear* - $t_{motor}(7)=4.9$, $p<.05$; *tune radio frequency* - $t_{visual}(7)=4.2$, $p<.05$; $t_{motor}(7)=12.1$, $p<.05$;). Unique task primitives were developed for each of these tasks using the mean focus-group ratings.

The baseline RNAV model required three new primitives, not contained in the TAWL, which were specific to approach-and land operations in commercial fixed-wing aircraft. These were: *visually acquire runway; manipulate yoke; manipulate pedals.* Table 1 presents the primitive workload values implemented in MIDAS based on the focus group estimates.

Table 1. Validated MIDAS workload primitives for the RNAV model

Task	Visual	Cognitive-		Motor		
		Spatial	Verbal	Fine	Gross	Voice
Push and Release	3.7	1.2		2.2		
Reach Object	3.7	1.2			2.6	
Say Message			5.3			4.5
Set Speed	3.9	3.0		5.0		
Set Flaps	4.0	2.2			4.3	
Set Gear	4.0	2.0			4.7	
Tune Radio	4.5	2.7		5.4		
Acquire Runway	5	3.7				
Manipulate Yoke	1.2	5.9		1.3	1.3	
Manipulate Pedals		5.1			3.2	

Note: These tasks do not contain an auditory component.

2.4 Output Validation

In the output validation phase, the model outputs of workload and dwell percentage from 10 monte carlo model runs were compared to empirical data from the existing literature. This phase was completed after all of the inputs into the HPM were modified based on the task trace and parameter input analyses described previously.

Method. The baseline RNAV model's predicted workload and percent dwell time (PDT) data were compared to empirical data from independent HITL simulations available in the literature. Statistical correlation tests were conducted to evaluate the goodness-of-fit between the model and HITL data. For all analyses, only the PF data are shown.

A survey of the literature was conducted to identify relevant HITL data sources from commercial pilots flying approach-and-land scenarios in a glass cockpit in either an actual flight test or a high-fidelity flight simulator. One HITL study was identified as a suitable comparison for the workload data [18]. This medium-fidelity HITL simulation was previously conducted for a different model validation effort, and as such was unique in that it provided workload estimates from three commercial pilots using the TAWL scale, for the three phases of flight modeled in the current baseline RNAV model. Three additional HITL studies [19,20,21] were identified as suitable comparisons for the visual fixation data. Each study included commercial pilots, flying Instrument Landing System (ILS) approach-and-land scenarios.

Results

Workload. Figure 4 (left) presents the overall workload as predicted by the MIDAS RNAV model and estimated by the pilots in the HITL simulation [18] for each of three phases of flight (descent, approach, and land). For the model data, overall workload was calculated as the mean of the individual workload channels (Visual, Auditory, Cognitive, and Motor) within each phase of flight from 10 monte carlo runs. The HITL data are the mean of the three subjects' subjective estimates of Overall Workload from a nominal baseline IMC scenario. As can be seen, the HPM and HITL data are positively correlated (r^2=.63). The model data tended to over-predict workload during the landing phase. It should be noted that the HPM simulation was a medium-fidelity simulation, and lacked the high-fidelity representations of the instrumentation and controls, which may have actually lead to an under-estimation of workload by the actual pilots in the HITL simulation.

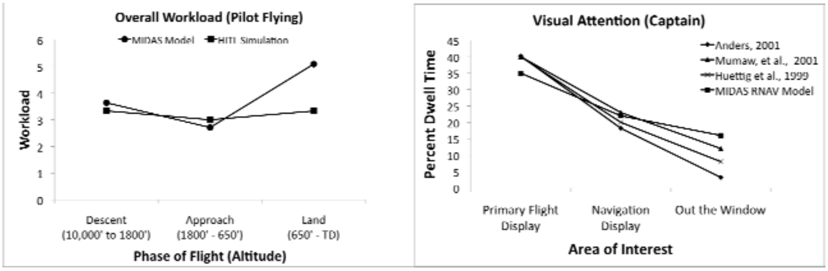


Fig. 4. Output validation: Overall workload (left), visual fixations (right)

Visual Fixations (Percent Dwell Time). Figure 4 (right) presents the model fit of PDT on three areas of interest: Primary Flight Display (PFD), Navigation (Nav) Display, and Out-the-Window (OTW). The model represents the PDT to each display for the entire scenario (10,000' to touchdown) averaged over 10 monte carlo runs. These data were compared to the three separate HITL data sets [19,20,21] from approximately the same phase of flight. There was a strong positive correlation (r^2 = .96) between the RNAV model PDT output and the average of the three HITL studies. This is evidence that both the model inputs and the SEEV process model, which guides visual attention, are related linearly.

3 Discussion and Summary

A HPM of commercial airline pilots conducting approach-and-land procedures was created using the MIDAS software following a methodical development and validation approach. The premise that guided the current work was that model validity is a process, not solely a single value at the conclusion of a model development effort. Valid inputs lead to valid outputs. Conducting only one of these validation processes may lead to invalid models. This is especially true as the complexity of the operational environment and tasks increase.

The pilot focus groups were instrumental in defining valid model inputs. The scenario-based cognitive walkthrough approach captured the context of operations well and enabled the pilots to easily identify tasks that depend on specific phases of flight, and augment the environmental considerations that are used to drive the model's performance.

The workload associated with the behavioral primitives was evaluated with some degree of success. This effort illustrated that MIDAS workload primitives, derived directly from the TAWL, were valid as evaluated by the focus-group pilots. Context-specific workload primitives were modified based on pilot input.

The model output correlated strongly with multiple independent human-in-the-loop simulation studies. These output validation results provide further evidence that the model inputs and the workload and SEEV process model are valid.

In summary, the methodical and comprehensive model validation effort presented in this paper illustrates a candidate process for developing and validating HPMs. This valid current-day RNAV model will next be extended to evaluate the impact of potential NextGen CSPO concepts.

Acknowledgements. The composition of this work was supported by the Federal Aviation Administration (FAA)/NASA Inter Agency Agreement DTFAWA-10-X-80005 Annex 5 (FAA POCs Dr. Tom McCloy, Mr. Dan Hershler; NASA POC Dr. David C. Foyle). The authors would like to thank Connie Socash, Dr. Christopher Wickens, Dr. Andrew M. Gacy, and Mala Gosakan from Alion Science and Technology (MAAD Operations) for their invaluable model development support and all reviewers for their insightful comments.

References

1. JPDO, Joint Planning and Development Office: Concept of operations for the next generation air transportation system. In: Joint Planning and Development Office (Ed.), vol. 3. JPDO, Washington, DC (2009)
2. Gore, B.F.: Chapter 32: Human Performance: Evaluating the Cognitive Aspects. In: Duffy, V. (ed.) Handbook of Digital Human Modeling. Taylor and Francis/CRC Press, NJ (2008)
3. Hooey, B.L., Foyle, D.C.: Advancing the state of the art of human performance models to improve aviation safety. In: Hooey, B.L., Foyle, D.C. (eds.) Human Performance Modeling in Aviation, pp. 321–349. CRC Press/Taylor & Francis, Boca Raton (2008)
4. Gore, B.F.: An emergent behavior model of complex human-system performance: An aviation surface related application. VDI Bericht 1675, 313–328 (2002)
5. Gore, B.F.: Man-machine integration design and analysis system (MIDAS) v5: Augmentations, motivations, and directions for aeronautics applications. In: Cacciabu, P.C., Hjalmdahl, M., Luedtke, A., Riccioli, C. (eds.) Human Modelling in Assisted Transportation. Springer, Heidelberg (2010)
6. Wickens, C.D., McCarley, J.M.: Applied Attention Theory. Taylor and Francis/CRC Press, Boca Raton (2008)

7. Gore, B.F., Hooey, B.L., Wickens, C.D., Scott-Nash, S.: A computational implementation of a human attention guiding mechanism in MIDAS v5. In: Duffy, V.G. (ed.) ICDHM 2009. LNCS, vol. 5620, pp. 237–246. Springer, Heidelberg (2009)
8. Arditi, A., Azueta, S.: Visualization of 2-D and 3-D aspects of human binocular vision. Paper Presented at the Society for Information Display International Symposium (1992)
9. Ericsson, K.A., Kintsch, W.: Long-term working memory. Psychological Review 102(2), 211–245 (1995)
10. Gugerty, L.: Evidence from a partial report task for forgetting in dynamic spatial memory. Human Factors 40(3), 498–508 (1998)
11. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology 47, 381–391 (1954)
12. Gore, B.F., Hooey, B.L., Haan, N., Socash, C., Gacy, M., Wickens, C.D., et al.: Evaluating NextGen Closely Spaced Parallel Operations Concepts with Human Performance Models. NASA Ames Research Center, Moffett Field (2011)
13. Keller, J.W., Leiden, K., Small, R.: Cognitive task analysis of commercial jet aircraft pilots during instrument approaches for baseline and synthetic vision displays. In: Foyle, D.C., Goodman, A., Hooey, B.L. (eds.) Aviation Safety Program Conference on Human Performance Modeling of Approach and Landing with Augmented Displays (NASA/CP-2003-212267). NASA Ames Research Center, Moffett Field (2003)
14. Gore, B.F., Hooey, B.L., Salud, E., Wickens, C.D., Sebok, A., Hutchins, S., Koenecke, C., Bzostek, J.: Identification Of Nextgen Air Traffic Control and Pilot Performance Parameters for Human Performance Model Development in the Transitional Airspace. In: NASA Final Report. ROA 2006 NRA # NNX08AE87A, SJSU, San Jose (2009)
15. Campbell, G.E., Bolton, A.E.: HBR validation: interpreting lessons learned from multiple academic disciplines, applied communities, and the AMBR project. In: Gluck, K.A., Pew, R.W. (eds.) Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation and Validation, pp. 365–395. Lawrence Erlbaum & Associates, New Jersey (2005)
16. McCracken, J.H., Aldrich, T.B.: Analysis of Selected LHX Mission Functions: Implications for Operator Workload and System Automation Goals (Technical note ASI 479-024-84(b)). Anacapa Sciences, Inc. (1984)
17. Hamilton, D.B., Bierbaum, C.R.: Operator Workload Predictions for the Revised AH-64A Workload Prediction Model: Volume I: Summary Report (AD-254 198). Anacapa Sciences, Inc., Alabama (1992)
18. Hooey, B.L., Foyle, D.C.: Aviation Safety Studies: Taxi Navigation Errors and Synthetic Vision System Operations. In: Hooey, B.L., Foyle, D.C. (eds.) Human Performance Modeling in Aviation, pp. 321–349. CRC Press/Taylor & Francis, Boca Raton (2008)
19. Anders, G.: Pilot's Attention Allocation during Approach and Landing: Eye- and Head-Tracking Research in an A330 Full Flight Simulator. In: Proceedings of the 11th International Symposium on Aviation Psychology, Columbus, OH, USA (2001)
20. Mumaw, R.J., Sarter, N., Wickens, C.D.: Analysis of pilots' monitoring and performance on an automated flight deck. In: Proceedings of the 11th International Symposium on Aviation Psychology. The Ohio State University, Columbus (2001)
21. Hüettig, G., Anders, G., Tautz, A.: Mode Awareness in a modern Glass Cockpit – Attention Allocation to Mode Information. In: Proceedings of the 10th International Symposium on Aviation Psychology, Columbus, OH, USA (1999)