Flight Deck Models of Workload and Multi-Tasking: An Overview of Validation

Christopher D. Wickens

Angelia Sebok

FLIGHT DECK MODELS OF WORKLOAD AND MULTI-TASKING: AN OVERVIEW OF VALIDATION

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

We review 24 computational modeling efforts of pilot multi-task performance and workload, to describe the manner in which they model three different aspects of pilot performance: the complexity of effort, the complexity of time management and the complexity of multiple resource interference. We then discuss the degree of validation of these models, and the validity of the context in which they are validated.

The next generation of the airspace program in the US (NextGen) calls for a variety of new concepts of operations as well as technologies, such as self separation, data-linked messages, closely spaced parallel operations and so forth (FAA, 2012). Naturally it is critical that such features preserve safety, even as they will increase efficiency. Traditionally, assessments of safety and efficiency of new technology and procedures is accomplished by human-in-the-loop simulation, yielding results based on response time, errors, workload, and more recently, situation awareness (Strybel et al., 2013). However it is also well understood that such HITL simulations can be extremely time consuming, and often lacking in statistical power because of a small sample size.

A complementary approach we advocate in this paper is the use of pilot or controller **computational models** of human performance that can be used to estimate the level of performance that will be achieved in certain conditions (Foyle & Hooey, 2008; Pew & Mavor, 1999; Gray, 2007). To the extent that *such models are valid* (an issue we address extensively in this paper), they can provide satisfactory predictions of performance in a fraction of the time, and at a fraction of the cost of full HITLs.

This paper describes one part of a project we performed for the FAA to establish the state of the art of computational models of pilot (e.g., flight deck) performance. What studies have been done? What aspects of performance have been modeled? And how valid are these models? In all, we identified 187 separate **modeling efforts**: typically a separate research paper that describes the application of a model to pilot performance. Many of these articles described the same **model architecture**, as applied to two or more different applications, or sets of PITL data. We then classified these efforts in terms of 13 different **aspects** of pilot performance, such as situation awareness, pilot error, pilot-automation interaction, and so forth. Details of each of these can be found in our full report (Wickens Sebok et al, 2013). However the current paper focuses only on two closely related aspects of pilot performance models: workload and multitasking. The reason for this restriction is twofold. (1). Workload and multitasking issues are of critical importance in the flight deck, as satellite navigation and improved sensors are transferring more responsibilities and tasks from ground to the fight deck. (2) These areas capture a long standing theoretical and practical interest of the first author, in their applications to aviation (e.g., Wickens Goh et al., 2003. Wickens, Sandry & Vidulich, 1983; Wickens & McCarley, 2008).

Model validation. It is our position that the best measure of model validation reflects the ability of a model to accurately predict performance (including measures like workload or situation awareness) across a **set** of flight conditions (e.g., NextGen [Ng] vs conventional [C] procedures, with advanced [A] versus older [O] equipment), so that the **differences** or variance between such conditions is accurately predicted. Such prediction, for the 4 condition case described above, is represented in figure 1, and is best captured by the product-moment correlation (r) between model predictions and P(Pilot)ITL data. While r describes the overall success of the model, the graphic scatter plot (e.g., figure 1) provides additional information concerning which conditions may be over- or under-predicted by the model. Furthermore, while the correlation coefficient may be the benchmark or gold standard of validation, other validation efforts short of this may still provide useful information. In the following, we do not discriminate between these different levels of validation, (but see Wickens, Sebok et al., 2013).

Workload and Multi-tasking. The concepts of workload and multi-tasking are closely related, both relating to the limits of the pilot's information processing capacity, but also distinct (Wickens & McCarley, 2008). Mental workload, generically relates to the relation between the total **demands** on that capacity imposed by single, or by multiple tasks, and the **availability** of cognitive resources to meet that demand. While higher workload may often diminish performance, it does not necessarily do so, if demands do not exceed operator capacity; hence effective measures of workload are often found, not in performance measures, but in physiological or subjective measures (Wickens & Yeh, 1988).

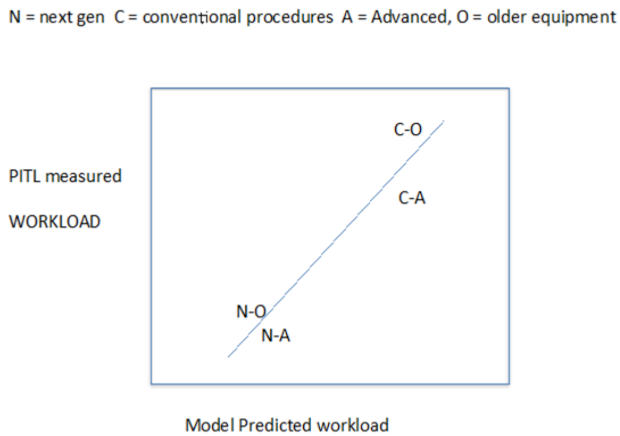

Figure 1: Correlational scatter plot representation of model validation of four predicted conditions. The regression line would suggest a high correlation (r >0.90) and hence strong validity.

In contrast, while performance will depend partially on the relationship between resources available and demanded (but only **when the latter exceeds the former),** there are other factors that may degrade performance, particularly when the latter depends on different (multiple) resources within human information processing. For example a pilot who must simultaneously activate two controls with the same hand may not experience high workload; but dual task performance of the tasks depending on the two controls will suffer (see Wickens Sebok et al, 2013).

Results: Workload and multi-tasking model validation.

Altogether 24 modeling efforts were identified that focused on either the workload or multitasking aspect (some of these may have also focused on other aspects as well), and most of these contained empirical PITL data in their write up (although these data were often not adequate to be considered true validation). We represent these efforts in the context of Table 1 which represents the multi-dimensional array of model types. Table 1 contains three columns. Each column is arrayed on a continuum of increasing model complexity or sophistication from bottom to top. The dimension of complexity are separate for single resource models of effort (left column) and of time (center column) and for multiple resource models (right column). However here we note that greater complexity typically requires greater sophistication, specialization and training of the model user; in a way that often inhibits the ubiquity of the model, even as it might also avail greater precision of model predictions. We also note that a given model may populate more than one column, being for example, simple on one column, but more complex on others.

Within each column, each number represents a unique modeling effort, whose identity can be found in part A of the reference list. Each effort (number) is in turn associated with two additional attributes. We have made an effort to classify the extent to which the modeling effort is **validated** or not, characterized by the underline of the ID number, and the extent to which we consider the modeling effort carried out on tasks or within a context that may be considered **high fidelity or validity to transport aircraft** (e.g., cockpit automation, use of line pilots in validation data) or not, characterized by the **bold face** coding for higher fidelity. Thus a modeling effort coded by both attributes (e.g., **67**) would be considered particularly valuable to the FAA.

Table 1: Three aspects of model complexity increasing from bottom to top. Each model effort is coded by a number, whose full citation is found in part A of the reference list. **Bold faced**: high fidelity. Underlined: validated.

		Demand-resouce interacion
		Weighted additive conflict be resources
		5 **14** 15 23 24
Computational model of cognitive complexity	Formal models: (ACT-R)	"VACP" (Overload/channel) (
1 11 **18**	1 5 12 16 17	**4** 15
		"VACP" summed over chann
	Simple queuing models	**4** 15
Effort (SWAT/TLX/Bedford, etc.)	10 22	"VACP" Table Lookup
Pilot assigned or SME assigned	Task overlap?	**2** **3** **4** 19 **20**
13	7 15	
	Time (TR/TA) TLAP	**MULTIPLE RESOURCES**
	11 21	
SINGLE	RESOURCES	

At the base/foundation of the table, we consider the simplest view of attention or processing capacity as a limited single resource. At this fundamental level, there are two different conceptions as to what this resource may be. On the left, it is considered to be a limited "pool" of processing effort (Kahneman, 1973), that has a physiological basis in brain metabolism (Parasuraman & Rizzo, 2007; Wickens, Hollands et al, 2013, chpt 11), and can be assessed by physiological measures such as heart rate variability, or by subjective rating scales. In the center column, the single limited resource may be considered **time**, and hence workload may be characterized by the ratio of time required to time available; and performance breakdowns related exclusively to the extent to which the latter former exceeds the latter. We now describe the increasing complexity(from bottom to top) within these columns in more detail.

Single Resources: Effort. Single resource models, where resources are based on the concept of effort may, on the one hand, simply have as inputs, pilot or SME assigned values of the effort of certain tasks (e.g., a SWAT or TLX rating). However these become somewhat circular to the extent that model validation is itself based upon a subjective workload assessment measure. More valuable, but more complex, are models in which predicted workload is computed on the basis of some objective computational algorithm of cognitive, perceptual or motor complexity (e.g., Boag et al, 2007) directly related to mental workload. For example in [18], a model of FMS complexity is derived, based on number of elements (modes) and interrelationship between them.

Single Resources: Time. When time is considered the single resource for which all tasks compete, there are many more modeling efforts that have varied in their level of complexity of the treatment of predicted time allocation between tasks that are not performed concurrently (see center column, table 1). At the simplest level, Parks & Boucek (11) have developed an aviation time-line-analysis procedure (TLAP) that simply tallies the total time demand of all tasks within an interval and divides this total by interval length to compute a predicted workload level. Such a technique can be made a bit more complex (and accurate) to the extent that amplified penalties are assigned proportional to the time that two tasks must make demands for the same (**overlapping**) period of time, such as when a pilot must lower a landing gear while attending to an ATC communication.

A layer of complexity is then added to the extent that models depend on formal queuing theory for their scheduling of tasks [22], and penalties are assigned for tasks that must wait until they are "served" by the single server queue (the pilot). Finally, at what we consider the highest level of complexity are those models that depend upon a formal well-validated (often in non-aviation domains) architecture, such as ACT-

R, or GOMS, or a derivative thereof. Such models are often applied in aviation to predict other aspects than workload (See Wickens Sebok et al, 2013); but they typically do include some time-management routine that will dictate the sequence in which tasks are performed, the period in which a task may be "neglected:, and hence an implicit measure of multiple task performance breakdown.

Multiple resources. While single resource models may be extended either in the direction of effort or time, the third dimension of complexity is to consider that resources are not single at all. Depicted in the right column, instead the pilot's information processing system contains separate resources such as visual versus auditory perception or vocal vs manual responses. To the extent that two tasks rely upon separate resources, they will be more successfully time shared (although will not necessarily lower workload). Thus the pilot will be more likely to hear an auditory warning signal if she is simultaneously reading a data link message (visual) than hearing an auditory ATC communication. The identity of these multiple resources is typically based on a fundamental model developed by Wickens (1984) that defines resources along three (now four) dichotomous dimensions (See Wickens Hollands et al, 2012 for a current version). Because these separate resources were originally assigned to four categories by model developers (visual, auditory, cognitive and psychomotor; Aldrich et al, 1989), we generically refer to this as the "VACP approach", even as the actual complexity of what defines resources is sometimes increased and may vary slightly between applications (see text below).

Fundamental to all such models is that tasks are assigned by a SME, model runner or "table lookup" to one or more resource types (e.g., comprehending ATC instructions will be considered an "A-C" task within the VACP model). Then within each resource a **demand level** is selected, often on a 1-7 scale. Some models stop at this point; others create a simple workload prediction by summing these values across channels, and at a still greater level of complexity some examine the extent to which any single channel is "overloaded" (e.g., cognitive value >5) by the sum of all task demands within that resource. Such overload predicts a multiple task performance breakdown. Finally, most faithful to the original multiple resource models are those that assign **weighted penalties** to task pairs to the extent that they share more dimensions **within the multiple resource space.** (Wickens, 2008). Thus for example two linguistic perceptual tasks will create more conflict (higher predicted penalty) than a linguistic and a spatial perceptual task.

Discussion: Validity, validation and complexity.

Tallying the two codes assigned in Table 1, we reach the conclusion that 15 of the 24 model efforts have been validated; but of those 15, only four appear to be carried out in what might be described as a high fidelity context. Correspondingly, of the 14 high fidelity studies, only 4 report careful validation.; and the reader should be aware that we have defined these two criteria liberally. If we were to restrict "high fidelity" to true NextGen concepts, and "validation" to use of correlations the number would be reduced substantially.

We also note the paucity of high fidelity validation toward the top of the table, where complex models predominate. Indeed two of these four only address cognitive complexity of component aspects of the flight deck. We account for this state of affairs in more detail in Wickens, Sebok et al, (2013), but note here the difficulty of accomplishing full validations of complex models, with the limited resources often made available for modeling efforts. However in this regard, we note and emphasize that *every modeling effort need not be validated*. Once a model **architecture** is validated in one context, greater faith can be held, that its un-validated predictions will nevertheless be accurate in a different context, hence allowing the great shortening of the time required to assess the viability of NG technology and procedures that we discussed at the outset of this paper.

References

A. Numbered references in table 1

[1] 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.*

[2] 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.

[3] Gore, B.F. & Corker, K.M., (2000b). Value of Human Performance Cognitive Predictions: *Human Factors and Ergonomics Society Annual Meeting Proceedings*. Santa Monica, CA: HFES.

[4] Gore, B. F., Hooey, B. L., Socash, C., et al.. (2011). *Evaluating NextGen closely spaced parallel operations concepts with human performance models* (HCSL, Trans.) HCSL Technical Report (HCSL-11-01). Moffett Field, CA: NASA Ames Research Center.

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

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

[8] Lyall, E.A. & Cooper, B., (1992). The Impact of Trends in Complexity in the Cockpit on Flying Skills and Aircraft Operation. *Human Factors Society Annual Meeting Proceedings*. Santa Monica, CA.

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

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

[11] Parks D. & Boucek ,G. workload prediction: challenges. In McMillan, G.R., Beevis, D et al.. *Applications of human performance models to system design* . New York City, NY: Plenum Press

[12] 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*. Ohio State.

[13] 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.

[14] 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.

[15] Sarno, K. & Wickens, C. (1995). The role of multiple resources in predicting time-sharing efficiency:. *International Journal of Aviation Psychology, 5*(1), 107-130.

[16] Schoelles. M.J., Gray, W.D. (2011). Cognitive Modeling as a Tool for Improving Runway Safety. *T Proceedings of the 16th International Symposium on Aviation Psychology*. Dayton, OH. 541-546.

[17] 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.

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

[19] See, J.E. & Vidulich, M.A. (1998). Computer Modeling of Operator Mental Workload and Situational Awareness in Simulated Air-to-Ground Combat:. *The International Journal of Aviation Psychology*. 8(4), 351-375.

[21] 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.

[22] 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

[23] Wickens, C.D., Harwood, K., et al.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. Pp.168-172.

[24] 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.

B. References cited in text.

Aldrich, T.B.,et al.. (1989) The development and application of models to predict operator workload during system design. In G. MacMillan et al (Eds.), *Applications of human performance models to system design* (pp. 65080). New York: Plenum.

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.

FAA (2012, March). *NextGen Implementation plan*. Washington D.C: Federal Aviation Administration.

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

Kahneman, D. (1973) *Attention & effort*. Englewood Cliffs N.J.: Prentice Hall.

Strybel, T, et al (2013) Measuring the impact of NextGEN Operating concepts on Situation Awareness & Workload *Int J. Aviation Psych. 23,* 1-26.

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

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W., & Talleur, D. A. (2003). Attentional models of multi-task pilot performance using advanced display technology. *Human Factors, 45*(3), 360-380.

Wickens, C. Hollands, J. Banbury S.& Paraasuraman R.(2012) aEngineering Psychology & Human Performqnxe 4th edition. Upper Saddle River N.J.: Pearson

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

Wickens, C. D., Sandry, D., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, output, and central processing. *Human Factors, 25*, 227-

Wickens Sebok, A. Keller, J. et al (et al. (2013). Modeling and Evaluating Pilot Performance in NextGen-. FAA final report DTFAWA-10-X-800. Annex 1.11

Yeh, Y. & Wickens, C.D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors* 30(1): 111-120.

Acknowledgments

This research was sponsored by grant #DTFAWA-10_X-800 from the Federal Aviation Administration to AlionScience, via NASA Ames Research Center. The authors gratefully acknowledge Dr David Foyle of NASA and Dr. Brian Gore of San Jose Research Institute for their scientific/technical oversight of this work; and of John Keller for his contributions to the research team. The opinions expressed in this report are solely those of the authors.