Journal of Artificial General Intelligence 2(2) 108-127, 2010 Submitted 2009-12-01; Revised 2010-11-08
10.2478/v10229-011-0012-6

Validating Computational Cognitive Process Models across Multiple Timescales

Christopher W. Myers CHRISTOPHER.MYERS.29@US.AF.MIL
Kevin A. Gluck KEVIN.GLUCK@US.AF.MIL
Glenn Gunzelmann GLENN.GUNZELMANN@US.AF.MIL
Air Force Research Laboratory

Michael Krusmark MICHAEL.KRUSMARK@MESA.AFMC.AF.MIL
L3 Technologies

Editor: Christian Lebiere, Cleotilde Gonzalez, and Walter Warwick

Abstract

Model comparison is vital to evaluating progress in the fields of artificial general intelligence (AGI) and cognitive architecture. As they mature, AGI and cognitive architectures will become increasingly capable of providing a single model that completes a multitude of tasks, some of which the model was not specifically engineered to perform. These models will be expected to operate for extended periods of time and serve functional roles in real-world contexts. Questions arise regarding how to evaluate such models appropriately, including issues pertaining to model comparison and validation. In this paper, we specifically address model validation across multiple levels of abstraction, using an existing computational process model of unmanned aerial vehicle basic maneuvering to illustrate the relationship between validity and timescales of analysis.

Keywords: model validation, evaluation, comparison, timescales, large-scale modeling

1. Introduction

There is an affinity among the high-level objectives being pursued by researchers studying artificial general intelligence and by researchers studying human cognitive architecture. Recent descriptions of the re-emergence in the artificial general intelligence (AGI) community of an interest in *general-purpose* intelligence capable of working in a variety of novel, unexpected situations, has included language that explicitly relates that goal to human-level intelligence. For instance, Wang (2009) describes the objective that artificial general intelligence be “comparable with that of the human mind” (p. 1) and Laird et al. (2009) call for an artificial functionality “able to approach the breadth and depth of human-level intelligence” (p. 1). These dual emphases on *general* functionality and being on par with *human* mind/intelligence align the objectives of AGI with the objectives of cognitive architecture.

On the cognitive architecture side, we also see language that bridges these communities. Newell’s (1980, 1990) descriptions of evaluation criteria, as well as the more recent treatment of that topic by Anderson and Lebiere (2003), strongly emphasized “flexible behavior” via computational universality, going so far as to identify this as the most important criterion on which cognitive architectures should be tested. Others have also emphasized breadth and generality in their descriptions of what cognitive architectures are intended to be, such as “a broad theory of human cognition” (Byrne, 2003), “a domain-generic computational cognitive model” (Sun, 2004), or “a software implementation of a general theory of intelligence” (Laird, 2008).

Given the significant overlap in missions and methodologies, it naturally is the case that these scientific communities share some common challenges. One shared challenge is the widely varying timescales and concomitant levels of complexity across which both natural and artificial intelligences behave (Simon, 1999). In the context of cognitive architecture, Newell (1990) referred to this as the timescale of human action, with times ranging from biological activity occurring over a period of microseconds to social activity occurring over a period of months (see Table 1). Computational process models that span these levels are a challenge for cognitive modeling (Anderson, 2002) as well as AGI (Laird et al., 2009). Indeed, as progress is made on AGI that is capable of acquiring knowledge and adapting over longer and longer periods of time, Newell’s entire timescale of human action becomes relevant to the artificial. Hence, a host of open issues in cognitive architecture associated with measurement, explanation, validation, and comparison across levels of analysis are also important for AGI.

Scale (sec)	Time Units	System	World (theory)
10^{7}	Months		
10^{6}	Weeks		Social Band
10^{5}	Days		
10^{4}	Hours	Task	
10^{3}	10 minutes	Task	Rational Band
10^{2}	Minutes	Task	
10^{1}	10 seconds	Unit task	
10^{0}	1 second	Operations	Cognitive Band
10^{-1}	100 milliseconds	Deliberate act	
10^{-2}	10 milliseconds	Neural circuit	
10^{-3}	1 millisecond	Neuron	Biological Band
10^{-4}	10 microseconds	Organelle	

Table 1. Newell’s Timescale of Human Activity (from Newell, 1990)

Model comparisons and challenges (such as the Dynamic Stocks and Flows Model Comparison Challenge; Lebiere, Gonzalez, & Warwick, this issue) are of great importance to developers of cognitive architecture and AGI systems, providing each with a means to measure progress. There are two general, yet complementary, approaches to model comparison—model-to-model comparison and model-to-referent comparison. In model-to-model comparison, two or more models are compared against each other and evaluated based on some predetermined metric, such as response times or task performance. In model-to-referent comparison, a model is compared against a referent (e.g., human data from a task of interest) and is evaluated by how closely the model data resemble the referent data. Indeed, model-to-referent comparison is often how members of the cognitive architecture community validate their models—comparing their model data to human data. Model comparison challenges, such as the Dynamics Stocks and Flow Challenge, use a combination of the model comparison approaches. This model comparison challenge determined the “best” model among a set of submitted models by examining how well data from each model predicted human performance from a referent data set.

In the current paper we adopt the model-to-referent approach for evaluating a model’s degree of validity first for its intended use, and then for an unintended use to address model generality. We demonstrate how questions of model validity increase in complexity when potential applications extend beyond purposes for which the model was originally developed. This raises important issues in the context of model comparisons and challenges, where determining which model is “better” is inextricably linked both to specific evaluation context(s) and to assumptions regarding the possible future uses for a model (i.e., model generality). We also discuss issues that arise when theorized cognitive mechanisms occur at timescales different from the knowledge, processes, or outcomes they predict. We demonstrate that Newell’s (1990) timescale of human action can serve as an organizing framework for evaluating, comparing, and contrasting the validity of cognitive models and AGI systems.

1.1 Computational Cognitive Modeling

Computational cognitive models can come in many forms, from a single equation (e.g., Bayes’ theorem) to systems of systems (e.g., ACT-R, EPIC, Soar, see Anderson, 2007; Kieras & Meyer, 1997; Wray & Jones, 2005, respectively). Growing out of the desire to develop computationally derived predictions of human behavior, scientists have heeded Newell’s (1973) call to stop playing 20 questions with nature and begin developing unified theories of cognition (Newell, 1990) through the integration of accumulated empirical knowledge and existing models (Gray, 2007). Cognitive architecture is one response to Newell’s call, and is intended to account for invariant aspects of human cognition (e.g., memory retrieval mechanisms, action selection, etc.). Cognitive architecture is typically instantiated as software, and serves as a foundation for the development of computational cognitive process models (Byrne, 2003; Gluck, 2010).

Much of the history of cognitive modeling has involved the understandable scientific strategy of isolating specific cognitive sub-systems through empirical studies in simple, abstract task contexts and then producing models that account for the empirical data. Examples include visual attention and search (Wolfe, 2007; Herd & O’Reilly, 2005), memory (Anderson & Schooler, 1991), problem solving (Newell & Simon, 1972), decision-making (Gonzalez, Lerch, & Lebiere, 2003; Lovett, 1998), and alertness (Gunzelmann, et al., 2009),.

As cognitive modeling continues to mature, it is becoming increasingly common to develop models that are capable of performing in more complex task environments, such as operating radar (Gray, Schoelles, & Myers, 2002; Taatgen & Lee, 2003), driving a car (Salvucci, 2006), flying an unmanned air vehicle (Gluck, Ball, & Krusmark, 2007) or a jet fighter (Jones, et al., 1999) or acting as a teammate (Ball et al., in press). These models depend on the integrated operation of a variety of cognitive processes (Gray, 2007), and quickly become large-scale systems-of-systems models. Large-scale models based on cognitive architecture contain hypotheses about underlying cognitive mechanisms at the architectural level, such as times associated with memory retrievals, as well as hypotheses about the knowledge and strategic approaches brought to bear on a task (Meyer & Kieras, 1997). These models may produce behavior across a wide range of times, from fractions of a second to days (Anderson, 2002; Newell, 1990; Simon, 1999).

How does a model developer determine the appropriate level of analysis for validating general, large-scale computational cognitive process models when different model components have effects on model behavior across a wide range of times? How is a model evaluated when used in contexts for which it was not originally designed? Finally, how can the cognitive modeling and AGI communities know which levels of analysis are appropriate for validating large-scale models? We take up these questions in the remainder of this paper, following the suggestion by Schoelles et al. (2006) that Newell’s (1990) timescale of human activity is an appropriate and useful framework for guiding model validation across multiple levels of analysis.

1.2 The Timescale of Human Activity

Allen Newell (1990) carved up human activity into timescales associated with biological, cognitive, rational, and social activities (see Table 1). Each of these bands captures approximately three orders of magnitude of the duration of various human activities. Newell suggested that cognitive architectures and models were best situated within the Cognitive Band.

Human behavior can be observed and measured at timescales of up to seven magnitudes greater than those occurring at the level where architectures and models are often situated (i.e., the Cognitive Band). Anderson (2002) challenged cognitive modeling to demonstrate that processes occurring at tens of milliseconds affect processes occurring over hours/days/months, and has implications for model validation. If the goal of a model is to predict human cognitive activity occurring within the Social Band (i.e., learning to fly a plane) using a model developed in a computational cognitive architecture situated within or below the Cognitive Band, then cognitive processes within the Social *and* Cognitive bands must be considered when validating the model (Anderson, 2002). Further, if measured behavior at the Social Band is hypothesized to result from processes scaling up from the Cognitive Band to the Social, then the model data from intervening levels of analysis must also be compared against referent data from the same levels (Schoelles et al., 2006).

The focus of our evaluation is a case study using Newell’s (1990) timescale of human activity as a frame for guiding and interpreting the validation of a cognitive model capable of performing flight maneuvers within an uninhabited air vehicle (UAV) synthetic task environment (STE). In the following sections, we introduce the task human participants and the model performed, followed by a description of the computational cognitive process model developed to account for expert pilot performance on basic flight maneuvers in the STE. We then describe the process undertaken to evaluate the validity of the model across two levels of human activity.

2. Model Validation in the Uninhabited Air Vehicle Basic Maneuvering Task

The STE used to collect data was developed for the Air Force Research Laboratory to provide an unclassified yet militarily relevant platform for basic research in human performance. The STE includes a realistic simulation of the flight dynamics of the Predator RQ-1A System 4 UAV, and three mission relevant tasks: basic maneuvering, reconnaissance, and landing. Basic maneuvering provided the task context for the research reported here.

2.1 The Basic Maneuvering Task

The basic maneuvering task (BMT) requires a pilot to make very precise, constant-rate changes in UAV airspeed, altitude and/or heading (Martin, Lyon, & Schreiber, 1998). It is an instrument flight task, with no out-of-cockpit view, requiring the pilot to attend to flight instruments to alter the UAV’s altitude, airspeed, and/or heading over the course of a 60 or 90 second trial (see Figure 1). The task consists of seven distinct maneuvers, and pilots attempt to minimize root-mean-squared deviation (RMSD) from ideal performance on each of the performance measures for each maneuver. Prior to beginning each maneuver is a 10-second lead-in, during which the operator is instructed to fly straight and level. At the end of this lead-in, the timed maneuver (either 60 or 90 seconds) begins, and the operator maneuvers the aircraft at a constant rate of change with regard to one or more of the three flight performance parameters (airspeed, altitude, and/or heading).

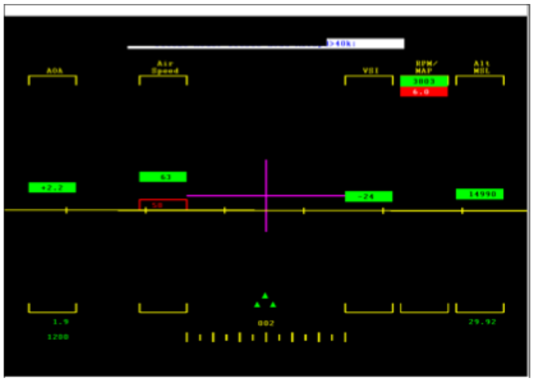

Figure 1. The UAV STE heads-up display.

The first three maneuvers require the operator to change one parameter while holding the other two constant. For example, in Maneuver 1 the task is to reduce airspeed by five knots at a constant rate of change over a 60-second trial while maintaining altitude and heading. Maneuvers progressively increase in complexity by requiring the operator to make constant rate changes along two and then three axes of flight. The seventh, and most difficult, maneuver requires changing all three parameters simultaneously over a 90-second trial: decrease altitude, increase airspeed, and change heading.

2.2 An Expert Model for the Basic Maneuvering Task

A computational cognitive process model of expert pilot performance in the basic maneuvering task was developed for use in contexts where a simulation of maneuver-level flying that is constrained by human cognitive limitations is functionally adequate, such as in a training simulation that requires a high cognitive fidelity representation of Predator maneuvering. The model was developed in ACT-R 5 (Anderson et al., 2004), which is a hybrid cognitive architecture that includes continuous processes that operate on symbolic knowledge. Symbolic knowledge is discrete and is divided into procedural knowledge that is implemented as IF-THEN rules (i.e., *production rules*) and declarative knowledge that represents retrievable facts (i.e., *chunks*).

The combination of continuous and symbolic processes situates the ACT-R architecture across the top half of the Biological Band (i.e., 10^{-2} sec; Table 1) and the lower end of the Cognitive Band (i.e., 10^{-1} sec). For instance, the declarative memory calculus within ACT-R results in latency effects ranging from 10s of milliseconds to 1 second (or even slightly more), and the default production cycle time is 50 ms. This provides a lower bound for which cognitive process models can be developed within ACT-R because it “abstracts away” from processes occurring at, and below, the 10^{-2} sec level of analysis. However, models developed in ACT-R can theoretically account for performance and phenomena occurring at the 10^{6} sec and the 10^{7} sec timescales. The theoretical claim and methodological approach within cognitive process modeling is that longer, more complex tasks are composed of collections of the atomic representations and processes at the lower timescales (Anderson, 2002). The basic maneuvering model is an example of this approach. The maximum duration of BMT maneuvers (60-90 sec.) falls within the Rational Band (i.e., 10^2 sec), providing an upper bound for applying Newell's timescale to BMT model evaluation.

2.2.1 The Control & Performance Concept for Instrument Flight

There is a prescribed strategy used by the United States Air Force for teaching instrument flight called the "control and performance concept" (USAF, 2000). The strategy is divided into two subtasks: control and crosscheck. The control subtask involves establishing appropriate control settings (i.e., stick and throttle positions) to achieve the desired performance result. Expert pilots have knowledge of the relationship between control instrument values (i.e., engine RPMs [power], pitch angle, and bank angle), and the performance characteristics they achieve in particular aircraft.

Importantly, changes in the performance of the aircraft unfold over the course of many seconds, whereas the impact on control instruments is more immediate. Thus, expert pilots first ensure that reasonable values are observed in the control instruments, and then initiate the crosscheck subtask, where the pilot verifies both that control settings are being maintained and that they are having the expected impact on performance (i.e., altitude, airspeed, and heading). Completing a maneuver requires repeatedly executing the crosscheck and control subtasks across 60 to 90 seconds. Hence, the control and performance concept, as a method for completing a maneuver, is best situated in the bottom Task level within the Rational Band (i.e., 10^2 sec; Table 1), whereas the control and crosscheck subtasks are best situated in the Unit Task level within the Cognitive Band (i.e., 10^1 sec).

The model implements the control and performance concept by executing the control subtask at the beginning of a maneuver, followed by the crosscheck subtask to assess performance and to ensure that control settings are being maintained. To effectively use the control and performance concept in the basic maneuvering task, the pilot must have the requisite control setting knowledge for various types of desired aircraft performance stored as declarative knowledge as well as the requisite procedural knowledge for adjusting the aircraft controls which are described in the following sections.

2.2.2 Declarative Knowledge for the Control and Performance Concept

The model's declarative knowledge is represented using four types of chunks: the goal chunk, crosscheck intent chunks, instrument chunks, and knowledge of appropriate control settings. The goal chunk contains knowledge and links to other knowledge that is needed to fly the UAV. The goal chunk contains knowledge that is organized into three categories of slots: maneuver knowledge (e.g., the duration of time passed in a trial), control knowledge (e.g., current and desired values for the control instruments), and performance knowledge (e.g., current and desired values for the performance instruments). Clearly this is much information, all of which is important to instrument flight; however, the production rules are designed so that only a few slots in the goal chunk are used for matching production rules at each procedural cycle.

Crosscheck-intent chunks are retrieved from declarative memory for moving attention from one instrument to the next, and determining whether the model performs the standard crosscheck or focuses on achieving appropriate control instrument settings for the maneuver. Retrievals are based on the instrument being attended, the maneuver, and the time into the maneuver. The model also has declarative knowledge of the instruments on the task display, including information about the location of the instrument and the significance of the value of that instrument for performing the task. Finally, the model has declarative knowledge of the control settings that are appropriate for executing the required maneuvers. This knowledge is crucial for establishing the correct settings at the start of a trial, following the lead-in period. Knowledge of the desired control instrument settings at given points in a scenario (e.g. 15 sec, 30 sec, 45 sec) is important for ensuring that performance objectives are being achieved.

2.2.3 Procedural Knowledge for the Control and Performance Concept

In order to succeed in the basic maneuvering task, the pilot must adjust the UAV stick and throttle to produce approximately the right rate of change in the performance instruments the moment a maneuver starts. Following the lead-in period of a maneuver, the model executes a series of productions to transition from the straight and level lead-in to performing the maneuver. This transition procedure is provided to the model since it is a model of expert performance and not a learning model. The execution of these productions is triggered by the perception of an auditory beep which occurs at the start of a trial following the lead-in period, via ACT-R's audition module, or by recognition that the lead-in period is nearing completion.

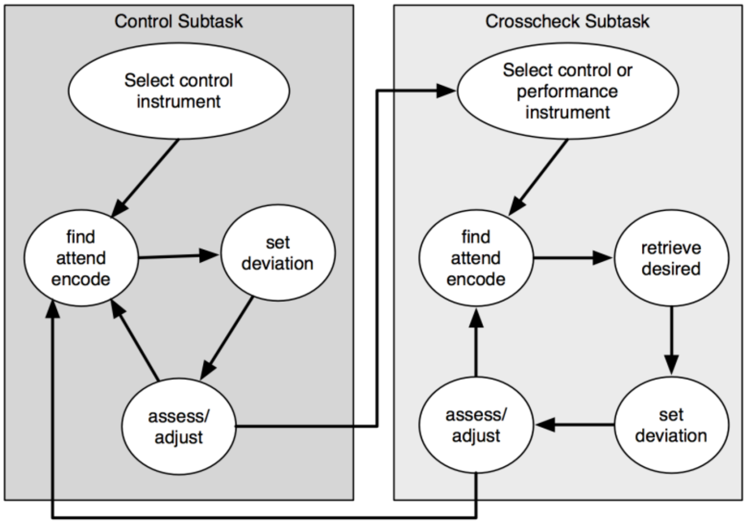

Figure 2. Steps in the control and crosscheck subtasks of the expert ACT-R model.

The model has separate sets of production rules that implement the control and crosscheck subtasks, each diagrammed in Figure 2. Establishing control begins with the selection of an instrument for which control needs to be established. This happens either at the beginning of a trial when the values of control instruments are first set, or whenever the value of a control instrument deviates beyond a threshold from the desired setting, which causes the model to focus on that instrument. To establish control, the model attends to control instruments rather than performance instruments at the beginning of a maneuver. If an adjustment to a control instrument is required, the model assesses the deviation between the current and retrieved values and then makes a control adjustment to the flight parameter. It then continues focusing on the same control instrument on the next production cycle to achieve the desired control settings. Once no other control instrument is in need of setting, the model sets its state to conduct a normal crosscheck.

The subtask process of crosschecking is very similar to the subtask process of establishing control (see Figure 2); however, crosschecks include attending to control *and* performance instruments. If the model attends to an instrument that deviates significantly from the desired value, it returns to the control subtask. Moderate deviations of instrument settings result in adjustments to the stick and/or throttle without exiting the crosscheck subtask.

2.2.4 Subsymbolic Processes

A variety of ACT-R parameters can be modified to influence model behavior at a subsymbolic level. In this model, however, parameters were not explicitly tuned to optimize the fit to human data. Default ACT-R values were used in cases where they were available, including (with values in parentheses): goal weight (1) and latency factor (1). Other relevant parameters were set to commonly used values, such as decay rate (.5), production utility noise (1), and activation noise (.25). These parameters directly affect the selection and retrieval times of declarative information (e.g., goal weight, latency factor, decay rate, and activation noise) as well as the selection between a set of unit tasks that are applicable given the same context (production utility noise).

3. Validation at the Task Level of the Rational Band

In the context of the UAV STE, each maneuver takes 60 or 90 seconds to complete. The control and performance concept falls within the Task level of the Rational Band; hence the model spans five levels of human activity (from 10^{-2} to 10^{2}). This section describes the validation effort at the model's highest band of human activity, the Rational Band (10^{2}).

3.1 Materials & Equipment

El Mar's Vision 2000 Eye-Tracking System was used to collect oculomotor data. The system estimates eye point of regard by recording horizontal and vertical eye position from the relative positions of corneal and pupil center reflections, and merging these data with a recording of the visual scene. El Mar's Fixation Analysis Software Technology was used to define eye fixations and generate data files that contain gaze sequences and times. Fixations were defined when the eye was stationary for a minimum of 167 milliseconds, and the eye was considered stationary below a velocity of 30 degrees per second. Fixations to specific flight instruments of the UAV STE were identified when fixations occurred within predefined regions for each instrument.

3.2 Participants

Human data were collected from seven aviation Subject Matter Experts (SMEs) at the Air Force Research Laboratory's Warfighter Readiness Research Division in Mesa, Arizona. Participants were active duty or reserve Air Force pilots with extensive experience in a variety of aircraft, but none had actual Predator UAV flying experience or training. All were mission qualified in Air Force operational aircraft, and all had commercial rated certification. The seven participants had an average of 3,818 hours flying operational aircraft.

3.3 Procedure

Participants completed each maneuver for a fixed number of trials that ranged from 12-24, depending on the difficulty of the maneuver. Each participant completed the maneuvers in order, starting with Maneuver 1 and ending with Maneuver 7. Success was defined as flying within the performance deviation criteria used by Schreiber et al (2002). The precise criteria for success are not critical here, but they were developed to be attainable yet challenging, even for expert pilots. We analyzed data from successful trials only because the model was developed as a performance model of skilled aircraft maneuvering. Hence, the appropriate comparison is between successful model and human trials.

3.4 Data Analysis & Results

The determination of overall task performance required aggregating data across airspeed, altitude, and heading deviation performance measures, which were each on different scales. To achieve this, the RMSD data for each performance measure was converted to a z score and the values were summed for each trial, providing a Mean Sum RMSD (z) score for each participant in each maneuver (49 scores total from seven participants on each of seven maneuvers). Mean Sum RMSD (z) scores were then averaged across maneuvers to obtain an average RMSD (z) for each participant. Those averages were used to compute a Grand Mean RMSD (z) score and a 95% Confidence Interval for participant performance. The grand mean and 95% CI are plotted in the left pane of Figure 3.

The model data are an average of 20 model runs in each maneuver. The model data were converted to z scores by a linear transformation, using the mean and standard deviation from the normalization of the RMSD’s in the SME data. Model data were aggregated in the same manner as the human data. The model data are plotted as a point prediction because we use exactly the same model for every run, without varying any of the knowledge or parameters that might be varied in order to account for individual differences. The model is a baseline representation of the performance of a single, highly competent UAV operator. There are stochastic characteristics (noise parameters) in ACT-R that result in variability in the model’s performance, so we ran it 20 times to get an average. This is not the same as simulating 20 different people doing the task. It is a simulation of the same person doing the task 20 times (without learning from one run to the next). The confidence intervals in the human data capture between-subjects variability. Due to the fact that we have just one model subject, it is inappropriate to plot confidence intervals for the point prediction.

The Task Level performance comparison indicates that the model flies the UAV at a level of proficiency equivalent to that of expert human pilots on average. If we de-aggregate down to the level of average performance on each maneuver, we see that the fit of the model to pilot performance does vary by maneuver (see Figure 3). Furthermore, across maneuvers, the model corresponds to human performance with an r^2 = .57 and a root mean squared scaled deviation (*RMSSD*; Schunn & Wallach, 2005) of 3.46, meaning that on average the model data deviate 3.46 standard errors from the SME data.

To get a better sense for how we should interpret these results, we used the jackknifing procedure to determine if the model data were similar to human data, which involved running the same goodness of fit measures for each of the human participants compared to the data from the other participants. We tested the fit of participant-1 (P1) to the data from P2-P7, then the fit of P2 to the data from P1, P3-P7, and so on. The average human fit is r^2 = .76 and *RMSSD* = 2.85. So the model’s fit to overall human performance is only slightly worse than the average individual human participant's fit to overall human performance. We interpret this as evidence that the model is a good approximation to expert performance in the basic maneuvering task.

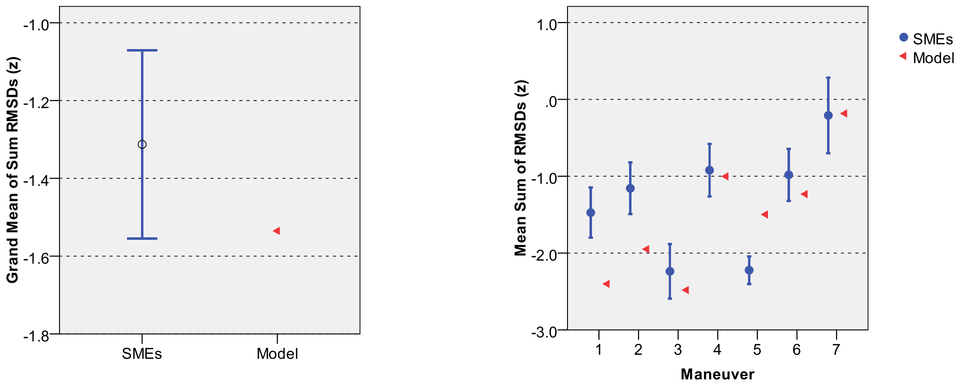

Figure 3. The left pane is the aggregate comparison of SME and model performance. The right pane is the comparison of SME and model performance by maneuver.

There are two things worth noting about the model data. First, the fact that it is a performance model and not a learning model could play a role in decreasing the fit to the human data. Note that model performance is better than the human data in the earlier maneuvers. Due to the fact that the SMEs progressed through the seven basic maneuvers in sequence, it would be reasonable to assume that rapid motor learning of the stick and throttle or adaptation to Predator-specific flight dynamics occurred within SMEs during Maneuver's 1 and 2 relative to Maneuver's 3 through 7. Maneuver 1 required SMEs to learn system specific stick pitch and throttle settings and resulting flight dynamics, while Maneuver 2 required additional learning of stick roll. This would explain the relatively large performance difference between SMEs and the model on Maneuver's 1 and 2. In fact, if we compute the fit using only data from Maneuver's 3 through 7, r^2 increases to .80 and *RMSSD* drops to 3.06.

Second, it is noteworthy that the model is sensitive to maneuver complexity, defined as the number of flight parameters, or axes (i.e., altitude, heading, and speed), that were supposed to be changing during the maneuver. Significant main effects of the number of axes were observed for both the model, $F(2,118) = 70.09$, $p < .001$, and SMEs, $F(2,449) = 37.87$, $p < .001$. For both the model and SMEs, performance was significantly better on one-axis maneuvers compared to two-axes maneuvers, $t(118) = 8.67$, $p < .001$ and $t(449) = 2.98$, $p < .01$, and on two-axes compared to three-axes maneuvers, $t(118) = 5.10$, $p < .001$, and $t(449) = 6.90$, $p < .001$, respectively. Thus, the model captures maneuver difficulty, even though it was not intentionally engineered to do so. These effects emerge naturally from the general design of the model.

3.5 Summary of Model Validation at the Task Level of the Rational Band

The results from the model validation effort situated at the Task level of the Rational Band demonstrate that the ACT-R model is a good approximation of overall task-level performance achieved by expert pilots, and does well on individual maneuvers. This serves as evidence to support the model's validity for use in application contexts in which a simulation of maneuver-level flying (constrained by the bounded cognition of humans) is functionally adequate, such as in a training event that requires a model of Predator maneuvering.

We might ask whether the model is valid for other purposes, such as for making predictions regarding the effects of changes in the training protocol on pilot performance. In this case, the answer is clear: the model is not valid for this purpose because it is a performance model, and not a learning model. Although the model does learn some declarative information in the process of completing a maneuver, it was not designed to take advantage of the full range of learning capabilities in the ACT-R architecture.

Another purpose of the model would be to fly an entire mission. However, the model is of basic maneuvers, and having the model complete entire missions would require changing the model by adding declarative and procedural knowledge to perform other mission relevant tasks along with basic maneuvers (e.g., Dimperio, Gunzelmann, & Harris, 2008). Another purpose of cognitive and AGI models is for evaluating how changes to a system interface affect task performance (Gray, John, & Atwood, 1993). For instance, proposed changes to the Predator's Heads-Up Display (HUD) could be evaluated through model predictions regarding whether the proposed changes have the desired effects, such as improved maneuver performance. Here it is less clear whether the model is valid for this purpose. This determination would likely depend on the nature of the change made to the HUD.

We might suppose in general, however, that the prediction validity of a model for evaluating system designs in complex visual environments like aircraft HUDs would depend very much on the construct validity of the underlying visual attention processes that guide information acquisition during task execution. This shifts the level of analysis for validation below the Task level, to the Unit Task level. Because the ACT-R architecture situates model development at the 10^{-2} sec and 10^{-1} sec levels of analysis and the basic maneuvering model spans five levels of analysis, it is not unreasonable to conclude that intermediate levels of analysis within the model should accurately replicate human cognitive processes, such as eye movements and movements of attention. In the next section, we re-evaluate the model by focusing on unit tasks, using fixation sequences (i.e., visual scans) as data. We compare visual scans from the SMEs with sequences of attention shifts generated by our expert model to investigate the extent to which the model may be valid for purposes that extend beyond the original modeling goal.

4. Validation at the Unit Task Level of the Cognitive Band

Sequences of fixations provide process-level information about human task performance, providing more detailed data regarding how the task is being done (Myers & Schoelles, 2005; Salvucci & Anderson, 2001). Such data are important if the goal is to create human-computer interfaces that optimize performance on critical tasks (Myers & Gray, 2010). Thus, assessing model validity for the predictive analysis of alternative interfaces requires that human performance be captured accurately at this lower, process-focused, Unit Task level of analysis. The UAV basic maneuvering model was not developed with this purpose in mind, and so this level of analysis was not explicitly considered during model development; however these processes had to be included for the model to complete the maneuver.

Technology reuse in new contexts is not a contrived circumstance; technology is often leveraged for uses that go beyond its original purpose. Indeed, it often is the case that additional possible applications are not apparent until after a technology is developed. Velcro© and Global Positioning Systems are good examples of base technologies that are used today in a wider range of applications than those for which they were originally intended. In the context of this paper, the applications for UAVs have expanded significantly from their original role in military reconnaissance to include strike operations in the military and border protection in domestic airspace.

Yet there are limits to the transferability of any technology. It is important to assess the appropriateness of a technology for any novel application. Our goal in this section is to evaluate the validity of the basic maneuvering model for system design analyses. We compare shifts of attention from the model and eye movements from SMEs to determine if the model scans flight instruments in a similar manner to SMEs. If it does, this would support the conclusion that the model could be appropriately applied to evaluate alternative HUD layouts or designs. If not, it would call into question the validity of the model for such a purpose.

4.1 Data Analysis & Results

Of the seven SMEs in the dataset, five had well-calibrated eye tracking data that could be used for detailed analysis. The previous ACT-R model runs had not been configured to provide data on shifts of attention on the display. With this addition in place, the ACT-R model was run 24 times on each maneuver. The order of UAV flight instruments attended by the model throughout each maneuver was saved to a data file for analysis. Like the validation of the model at the task level, only successful trials from SMEs and model runs were analyzed.

To assess the degree with which SMEs followed the control and performance strategy, specific flight instruments were coded as either control or performance instruments based on the Air Force flight-training manual (USAF, 2000). Recoding instruments as “control” or “performance” focused the analyses at the appropriate level of abstraction for determining reliance on the control and performance strategy and away from specific and idiosyncratic approaches to encoding flight instruments.

Because the UAV basic maneuvering task is a dynamic, interactive, complex task environment, behavior within the environment changes the state of the environment, which in turn can influence behavior. Consequently, the first 10 seconds leading into the maneuver (i.e., the lead-in) and the first 15 seconds of the maneuver (i.e., the first leg) were analyzed to minimize differences in task states across individual participants. Furthermore, only data from the easiest and most difficult maneuvers (i.e., Maneuver 1 and Maneuver 7, respectively) were analyzed to maximize any differences between model and human scanning strategies as a function of task difficulty.

There were 90 visual scans from Maneuver 1, equally divided between the lead-in and first leg of the maneuver. There were 118 scans in Maneuver 7, also equally divided between the lead-in and first leg of the maneuver. The ACT-R model contributed 22 scans to the lead-in and the first leg of Maneuver 1 and 20 to the lead-in and first leg of Maneuver 7. (Six trials were omitted because they did not result in trial success.) The remaining scans were divided between five SMEs (see Table 2 for a breakdown of the number of scans contributed from the model and each SME). There are fewer scans from the SMEs relative to the ACT-R model due to difficulties associated with eye tracking data loss and poor calibration to the eye tracking system.

To compare model scans to human scans, two steps were pursued. First, the similarities between pairs of scans were computed, followed by a second step, which grouped scans according to their similarity. To compute similarities between pairs of scans, the Levenshtein (1966) sequence alignment algorithm was applied. This algorithm bases similarity on a computation of the minimum number of edits (e.g., insertions, deletions and replacements) necessary to change one scan into another.

The standard Levenshtein algorithm can be biased when sequences vary in length. However, the distance value can be normalized to control for differences in lengths of compared visual scans providing a normalized similarity index (NSI, Myers & Gray, 2010). NSI controls for length by dividing the number of edits to change one scan into another with the number of fixations from the longest of the two scans. The NSI metric represents the maximum similarity between two visual scans.

UAV	Maneuver 1		Maneuver 7	
Operator	Lead-in	First Leg	Lead-in	First Leg
ACT-R	22	22	20	20
SME-1	2	2	5	5
SME-2	5	5	19	19
SME-3	4	4	1	1
SME-4	5	5	5	5
SME-5	7	7	9	9
Total	45	45	59	59

Table 2. Number of visual scans obtained from human and model AVOs by maneuver and leg. The lead-in is the first 10 seconds of a maneuver, and the first-leg follows the lead-in for 15 seconds.

The second step used the NSI values to determine *scanning strategies* by grouping scans according to their similarity. This was done by applying principle components analysis (PCA) on the computed NSIs. The number of components was determined by a rotated eigenvalue greater than one, where components with an eigenvalue less than one were assumed to be contributing little to the explanation of variance and were excluded from the analyses. Each component with an eigenvalue greater than one can be considered a general strategy and each scan within a component is a specific instantiation of the strategy. A limitation of PCA is the difficulty in providing descriptions of the components. *K*-means clustering was used in conjunction with PCA to ease the burden of component description[1].

ProtoMatch software (Myers & Schoelles, 2005) was used to compute the NSI values. Analyses were conducted for Maneuver 1, Lead-in (M1-L), Maneuver 1, Leg-1 (M1-1), Maneuver 7, Lead-in (M7-L), and Maneuver 7, Leg-1 (M7-1). NSI values were obtained for each sequence when compared against all other sequences within the same leg from the same maneuver. There were 990 NSI values from M1-L (M_{NSI} = 0.46) and M1-1 (M_{NSI} = 0.40). There were 1,711 NSI values from M7-L (M_{NSI} = 0.46) and M7-1 (M_{NSI} = 0.38).

For Maneuver 1, the PCA and *k*-means cluster analyses resulted in three components that accounted for 93.5% of the variance between NSI values from the lead-in, and four components that accounted for 89.1% of the variance between NSI values from leg one. The analyses for Maneuver 7 resulted in four components that accounted for 92% of the variance between NSI values from the lead-in, and four components that accounted for 90.4% of the variance between NSI values from leg one. The small number of clusters superficially suggests relative consistency in the scanning strategies used across human and model participants for each leg of each maneuver, and demonstrates visual scan similarity among participants.

Table 3 presents the results of the clustering analysis. In all cases but one, attention shifts from the model grouped separately from visual scans from the expert pilots. There is a clear tendency for a higher proportion of model attention shifts to control instruments than the human

[1] Ding and He (2004) proved that components from PCA are the continuous solutions to the discrete cluster membership associated with K-means clustering results. Consequently, using K-means clustering to aid in interpreting results from PCA will not be a factor in erroneous descriptions/interpretations of components/clusters.

participants. In addition, the model produces the same number of attention shifts compared to clusters containing human scans on average, and demonstrates that cluster solutions were not solely based on sequence length.

The finding that the clustering algorithm produced clusters that almost completely separated the human visual scans from the model scans raises important questions regarding the Unit Task level validity of this model. It suggests that one should be skeptical of using this model to evaluate changes to HUD design or other features of the UAV interface. The lack of correspondence suggests that there are real differences between the model and the human participants with regard to the composition of subtasks being used to execute the control and performance concept.

				Mean Root Mean Squared Deviation			Contributed visual scans		Percent of fixations on control items
Cluster		# of Scans	Mean Scan Length	Altitude	Airspeed	Heading	Human	Model	Mean% (*SD%*)
M1-L	A	4	5.25	16.48	0.82	0.03	4	--	53% (*6.5%*)
	B	19	13.68	15.37	1.37	0.22	19	--	52% (*9.1%*)
	C	22	13.91	8.6	0.92	0.00	--	22	100% (*0.0%*)
M1-1	D	13	25.31	14.79	1.19	0.83	13	--	41% (*8.8%*)
	E	11	20.82	12.37	1.12	0.19	7	4	61% (*26.9%*)
	F	18	20.22	8.9	0.93	0.00	--	18	84% (*9.2%*)
	G	3	12	19.82	1.69	0.00	3	--	47% (*8.2%*)
M7-L	H	18	8.83	12.79	1.3	5.4	18	--	63% (*11.8%*)
	I	18	13.89	16.41	1.59	4.89	18	--	59% (*9.6%*)
	J	3	5	10.85	1.52	3.9	3	--	92% (*14.4%*)
	K	20	14.25	14.27	1.4	4.73	--	20	100% (*0.0%*)
M7-1	L	20	20.45	14.43	1.4	4.73	--	20	99% (*1.2%*)
	M	14	15.43	14.22	1.4	5.08	14	--	57% (*8.2%*)
	N	5	12.2	10.07	1.49	5.62	5	--	62% (*9.7%*)
	O	20	20.25	15.43	1.47	4.88	20	--	61% (*6.3%*)

Table 3. Descriptive statistics associated with cluster solutions of NSI values from the lead in and first leg of maneuvers 1 and 7. Mean RMSD scores are the average root mean squared deviation between participants' altitude, airspeed, and heading and the desired settings for each flight parameter.

4.2 Summary of Model Validation at the Unit Task Level of the Cognitive Band

The results from the model validation effort at the Unit Task level are not nearly as promising as Task level results. The results revealed that the model did not attend to control instruments in the UAV basic maneuvering task in either the same sequence or in the same proportion relative to expert human pilots. Sequences of attention shifts from the model were clustered with human visual scans in only M1-1, but only to a limited extent even in that case. Furthermore, clusters did not represent individuals; instead, each cluster contained scans from multiple individuals. This is clear, as at most there were four clusters yet there were five SMEs and the model.

The findings lead us to conclude that scans produced by different SMEs were more similar to each other than they were to sequences of attention shifts produced by the ACT-R model. With respect to use of the control and performance concept, a large proportion of control instruments fixated during the lead-in and first leg of a mission demonstrate the adoption of the concept.

Inspection of Table 3 reveals that clusters dominated by sequences from the model had proportions of fixated control instruments near 90%, and clusters dominated by SME visual scans had proportions near 50%. Based on these results we would not be confident using the UAV basic maneuvering ACT-R model to provide predictions associated with the evaluation of new HUD layouts or designs for the UAV.

A potential limitation of the Unit Task level evaluation is that the comparison was between sequences of items attended by the model and sequences of items fixated by the SMEs. Research has demonstrated that shifts of attention can occur without saccades; however, there is a close tie between the two, as saccades tend to follow shifts of attention and are incapable of being executed to locations orthogonal to the location of attention (Kowler et al., 1995). A possible effect on the data is the shifts of attention without including the costs of eye movements could possibly lead to substantially more items attended by the model than fixated by the SMEs; however, the numbers of items attended by the model were not substantially different from items fixated by SMEs (see Mean Scan Length in Table 3).

5. Discussion

Creating a capacity for generally intelligent systems to complete tasks whose times range from tens of seconds to tens of months is desirable for AGI system developers and developers of computational process models grounded in cognitive architecture. Even more common ground is found in the desired ability for developed systems to successfully generalize beyond the task contexts originally intended by developers. Results from the current analyses provide insight into just some of the challenges associated with cognitive systems that can produce behavior across multiple timescales and their operation within unintended contexts.

We presented Newell's timescale of human activity as a guide to evaluating cognitive architecture and AGI across multiple levels of analysis. Based on the results, we are confident that the ACT-R cognitive model reported here would perform at the appropriate effectiveness for mimicking human behavior when performing UAV maneuvers. However, we are less confident that the model would be appropriate for the purpose of evaluating UAV HUDs.

When evaluating computational cognitive process models, it is often the case that models are developed at a timescale below the timescale of dependent variables collected from human participants to provide a referent. Consequently, it is difficult to evaluate cognitive processes across the intervening levels. As new data collection techniques become less cumbersome, such as analyzing eye movement sequences, we can begin to evaluate the intervening levels. Although we report a case study of model evaluation that moved down Newell's timescale of human activity, from 10^2 to 10^1, model evaluation can also move up the timescales. For example, if the basic maneuvering model were to be provided knowledge for completing whole missions, processes associated with situated action, such as when to execute different maneuvers or maintain situation awareness, become important processes for evaluation at higher levels of analysis (e.g., 10^3-10^4 in Table 1). Consequently, we conclude that Newell's timescale of human activity provides a useful organizing framework for evaluating cognitive models and AGI.

Given the results from the presented case study and developers' desiderata of model reuse, expansion, and integration with other models, how do developers determine the capabilities of their model when the model is being used for a different purpose, within a new context, or when models are being compared against each other? In the following sections we elaborate on these concerns.

5.1 Apparent Validity Disjunction across Timescales

How is it that a single model can vary in its apparent validity across timescales? Further, what can a developer deduce from validity disjunction across timescales? Before addressing these questions we must first consider validity and what a developer can conclude based on outcomes from validation exercises. If a single validation test is conducted on a model and model data fall within confidence intervals derived from a referent, should a developer conclude that the model is valid? Indeed, developers typically draw this conclusion on the basis of such results. However, what if the model data fall outside of confidence intervals of referent data, but are still correlated to referent data? In this situation a developer would be less confident in accurate predictions from the model, but more confident that the model is capturing potentially interesting variation within the referent data. Consequently, we would argue that labeling the model as either valid or invalid is not warranted. This is true in the conflicting evidence case just described and also more broadly in all evaluations of validity. Assessing the validity of a model, architecture, or system is an evidence accumulation process, with conclusions about validity best thought of as existing on a continuum. Recasting validity as a matter of degree of confidence regarding usefulness for a specific purpose, rather than as a globally relevant binary state, seems to us to be an appropriate way for researchers to orient their thinking and methods when evaluating cognitive architecture and AGI.

Based on the case study presented above, one interpretation of the differences in goodness-of-fit results between our model and referent human visual scan data at the Unit Task level is that there are likely many different sequences of behavior that can result in successful maneuver performance. In the case of the ACT-R model, the behavioral sequences being produced at the Unit Task level of analysis did not adequately capture the human approach to beginning a maneuver, nor the variety of approaches. However, the approach used by the model at the Unit Task level resulted in very similar performance outcomes to the approaches used by the SMEs. Indeed, this illustrates a persistent issue associated with model development, where an infinite number of models could account for a given data set. It may also be the case that expert pilots have visual scan strategies that illustrate a more sophisticated strategy than the textbook description of the Control and Performance Concept. Accumulated expertise may lead to a more complex interleaving of control and crosscheck subtasks, which would also explain why expert human pilots tended to have proportionally fewer fixations on control instruments than the model.

Developers typically have an intended use in mind when they begin work on a model. The intended use informs the developers which level of analysis should be the focus of the development effort. Further, when developing models to operate at a specific level of analysis, the precision of modeled processes below the specified level becomes less important to the developer. Indeed, this is the point of developing a model; it allows the developer to concentrate on precisely modeling processes that produce the range of behaviors of interest while abstracting away from finer-grained details that are not of interest. This is central to the approach of cognitive architecture where cognitive mechanisms are the focus, rather than neural-level processes. However, as models become more general in their intended use they must also be capable of performing tasks across multiple levels of abstraction. Hence, developers will have to pay increasing attention to the precision of mechanisms across multiple levels of analysis as model generality and complexity increase.

5.2 The Importance of Newell's Timescales as System Complexity Increases

Model comparison, such as the Dynamic Stocks and Flows Modeling Challenge, provides a useful technique for estimating progress. As developers of AGI and cognitive architecture move from building models of single tasks to building systems capable of learning and operating across a wide range of tasks and contexts for long periods of time, model precision will increase across multiple levels of analysis. Hence, generally intelligent agents will be required to operate across many of the different levels described by Newell, and will need to be capable of producing valid behavior within each level. Further, as research on the connection between the brain and cognitive activity is accumulated, it is likely that cognitive scientists interested in building a unified theory of human cognition will begin to integrate computational and mathematical models of biological processes that affect processes at higher levels of analysis. Indeed, this has already begun within the ACT-R architecture (Anderson, 2007; Gunzelmann et al., 2009).

We suggest that, first and foremost, validation efforts should be guided by models' intended uses. However, there will need to be specified approaches to measure progress toward attaining the goal of increased task generality as more developers embrace the goal of developing general, large-scale cognitive systems intended for uses that span multiple contexts and timescales. The question then becomes, how do we compare two models capable of operating across multiple timescales within multiple contexts? (To some, the sheer thought of attaining this goal suggests a valid cognitive system–if it can operate at such a large scale it must be doing something right!) In instances where the mere capacity to perform the task is insufficient to distinguish between models, referent data becomes a tool for comparing the models across the different levels of analysis. Within the domain of cognitive architecture, human behavior becomes the gold standard referent, and the number of levels with which a large scale model accurately replicates human referent data helps to place the model along the validation continuum. For example, models submitted as part of the Dynamics Stocks and Flows Modeling Challenge could be evaluated across a number of Newell's levels (e.g., 10^0, 10^1, and 10^2), along with models' performance on transfer tasks. The case study presented here demonstrated the use of human visual scans as referent data for the Unit Task level of analysis.

An important goal in cognitive architecture is for developers to stop being modelers and to become architects. This entails that generally intelligent agents based on cognitive architecture have the capability of being "set loose" in an environment with instructions on how to perform a multitude of tasks and produce learning and performance patterns that are in line with humans performing the same tasks, given the same instructions. This goal for cognitive architecture represents an example of an artificial general intelligence system that incorporates cognitive limitations inherent to human cognitive processing. Evaluating the robustness of such systems will require the capability for developers to perform model-to-referent comparisons across a range of levels within Newell's timescale of human action, and the same is true of AGI systems. The case study presented here provides a limited example of performing such model comparisons as well as highlighting issues of validity interpretation that will likely arise as AGI and cognitive architecture increase in generality.

Acknowledgements

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This research was sponsored by the U.S. Air Force Research Laboratory, 711th Human Performance Wing, Human Effectiveness

Directorate, Warfighter Readiness Research Division. The UAV model and fit to human task performance data were presented at the 5th International Conference on Cognitive Modeling in Bamberg, Germany.

References

Anderson, J. R. (2002). Spanning seven orders of magnitude: a challenge for cognitive modeling. *Cognitive Science, 26,* 85-112.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review, 111*(4), 1036-1060.

Anderson, J. R., & Lebiere, C. (2003). The Newell test for a theory of cognition. *Behavioral and Brain Sciences, 26,* 587-637.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. Psychological Science, 2, 396-408.

Ball, J. T., Myers, C. W., Heiberg, A., Cooke, N. J., Matessa, M., Freiman, M., et al. (under review). The Synthetic Teammate Project. *Computational and Mathematical Organization Theory*.

Byrne, M. D. (2003). Cognitive Architecture. In J. Jacko & A. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (pp. 97-117). Mahwah, NJ: Lawrence Erlbaum.

Dimperio, E., Gunzelmann, G., & Harris, J. (2008). An initial evaluation of a cognitive model of UAV reconnaissance. In J. Hansberger (Ed.), P*roceedings of the Seventeenth Conference on Behavior Representation in Modeling and Simulation* (pp. 165-173). Orlando, FL: Simulation Interoperability Standards Organization.

Ding, C., & He, X. (2004). *K-means Clustering via Principal Component Analysis.* Paper presented at the 21st International Conference on Machine Learning, Banff, CA.

Gluck, K. A. (2010). Cognitive architectures for human factors in aviation. In E. Salas & D. Maurino (Eds.) *Human Factors in Aviation, 2nd Edition* (pp. 375-400). New York, NY: Elsevier.

Gluck, K. A., Ball, J. T., & Krusmark, M. A. (2007). Cognitive control in a computational model of the Predator pilot. In W. D. Gray (Ed.), *Integrated models of cognitive systems* (pp. 13-28). New York, NY: Oxford University Press.

Gluck, K. A., Ball, J. T., Krusmark, M. A., Rodgers, S. M., & Purtee, M. D. (2003). *A computational process model of basic aircraft maneuvering.* Paper presented at the 5th International Conference on Cognitive Modeling, Universitats-Verlag, Bamberg.

Gonzalez, C., Lerch, F. J., & Lebiere, C. (2003). Instance-based learning in real-time dynamic decision making. *Cognitive Science, 27*(4), 591-635.

Gray, W. D. (Ed.). (2007). *Integrated Models of Cognitive Systems*. Oxford: OUP.

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating GOMS for predicting and explaining real-world task performance. Human Computer Interaction., 8(3), 237-309.

Gray, W. D., Schoelles, M., & Myers, C. W. (2002). *Computational cognitive models ISO ecologically optimal strategies.* Paper presented at the 46th Annual Conference of the Human Factors & Ergonomics Society, Baltimore, MD.

Gunzelmann, G. F., Gross, J. B., Gluck, K. A., & Dinges, D. F. (2009). Sleep Deprivation and Sustained Attention Performance: Integrating Mathematical and Cognitive Modeling. *Cognitive Science*.

Herd, S.A. & O'Reilly, R.C. (2005). Serial visual search from a parallel model. *Vision Research, 45,* 2987-2992.

Jones, R. M., Laird, J. E., Nielsen P. E., Coulter, K., Kenny, P., and Koss, F. Automated Intelligent Pilots for Combat Flight Simulation, AI Magazine , Spring 1999, Vol. 20, No. 1, pp. 27-42.

Kieras, D., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12*, 391-438.

Kowler, E., Anderson, E., Dosher, B., Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35*(13), 1897-1916.

Laird, J. E. (2008). *Extending the Soar cognitive architecture.* Paper presented at the 1st Artificial General Intelligence Conference, Memphis, TN.

Laird, J. E., Wray, R. E., Marinier, R. P., & Langley, P. (2009). *Claims and challenges in evaluating human-level intelligent systems.* Paper presented at the 2nd Artificial General Intelligence Conference, Arlington, VA.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10,* 707-710.

Lovett, M. C. (1998). *Choice.* In J. R. Anderson & C. Lebiere (Eds.) *The atomic components of thought,* 255-296. Mahwah, NJ: Erlbaum.

Martin, E., Lyon, D. R., & Schreiber, B. T. (1998). *Designing synthetic tasks for human factors research: An application to uninhabited air vehicles.* Paper presented at the Human Factors and Ergonomics Society.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive control processes and human multiple-task performance: Part 2. Accounts of Psychological Refractory-Period Phenomena. *Psychological Review, 104,* 749-791.

Myers, C. W., & Gray, W. D. (2010). Visual scan adaptation during repeated visual search. *Journal of Vision, 10*(8): 4, 1-14; doi:10.1167/10.8.

Myers, C. W., & Schoelles, M. (2005). ProtoMatch: A tool for analyzing high-density, sequential eye gaze and cursor protocols. *Behavior Research Methods, 37*(2), 256-270.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283-308). New York: Academic Press.

Newell, A. (1980). Physical symbol system. *Cognitive Science, 4,* 135-183.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H.A. (1972). Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall.

Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. *Human Factors, 48*, 362-380.

Salvucci, D. D., & Anderson, J. R. (2001). Automated Eye-Movement Protocol Analysis., *Human-Computer Interaction* (Vol. 16, pp. 39-86): Lawrence Erlbaum Associates.

Schoelles, M. J., Neth, H., Myers, C. W., & Gray, W. D. (2006). *Steps toward integrated models of cognitive systems: A level-of-analysis approach to comparing human performance to model predictions in a complex task environment.* Paper presented at the 28th Annual Conference of the Cognitive Science Society, Vancouver, BC, CA.

Schreiber, B. T., Lyon, D. R., Martin, E., & Confer, H. A. (2002). *Impact of prior flight experience on learning Predator UAV operator skills* Mesa, AZ: Air Force Research Laboratory, Warfighter Training Research Divisiono. Document Number)

Schunn, C. D., & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), *Psychologie der Kognition: Reden and Vortrage anlasslich der Emeritierung von Werner Tack* (pp. 115-154). Saarbrueken, Germany: University of Saaarland Press.

Simon, H. A. (1999). *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Sun, R. (2004). Desiderata for cognitive architectures. *Philosophical Psychology, 17*(3), 341-373.

Taatgen, N. A., & Lee, F. J. (2003). Production compilation: A simple mechanism to model complex skill acquisition. *Human Factors, 45*(61-76).

USAF. (2000). *Air Force Manual on Instrument Flight*. Retrieved. from.

Wang, P. (2009). Editorial: What makes JAGI special. *Journal of Artificial General Intelligence*, 1-2.

Wolfe, J. M. (2007). Guided Search 4.0: Current Progress with a model of visual search. In W. Gray (Ed.), Integrated Models of Cognitive Systems (pp. 99-119). New York: Oxford.

Wray, R. E., & Jones, R. M. (2005). An introduction to Soar as an agent architecture. In R. Sun (Ed.), *Cognition and Multi-agent Interaction: From Cognitive Modeling to Social Simulation* (pp. 53-78): Cambridge University Press.