Proceedings of the 2004 Winter Simulation Conference
R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

VALIDATION AND CALIBRATION OF HUMAN PERFORMANCE MODELS TO SUPPORT SIMULATION-BASED ACQUISITION

Floyd Glenn
Kelly Neville
James Stokes
Joan Ryder

1035 Virginia Drive
Suite 300
CHI Systems, Inc.
Ft. Washington, PA 19034, U.S.A.

ABSTRACT

We present a methodology under development for calibration and validation of human performance models in support of simulation-based acquisition processes — a human performance modeling validation program. We describe a conceptual framework based on an investigation of the characteristics of a wide variety of performance modeling frameworks and application domains. We offer initial taxonomies of model actions and empirical performance actions that will support the necessary mappings between model predictions and empirical observations for all models over the full range of required representation detail, and across all stages of the acquisition process in order to establish a solid analytic basis for calibration and validation of SBA processes and human performance modeling frameworks. We describe methods for specifying performance measures so that for any given design decision, performance measures captured using a model can be mapped to performance measures obtained during live test and evaluation.

1 INTRODUCTION

A variety of tools and methods are currently available for the construction of Human Performance Models (HPMs) in complex system and task environments (e.g., Pew and-Mavor 1998). Among other purposes, these models are being considered for use in simulation-based acquisition (SBA), providing tools for the interim evaluation of system performance at many intermediate stages in the acquisition process. Such HPM models can be used to represent human components of the system to be acquired, and/or human elements of the environment in which such a system would be employed. The various modeling tools offer somewhat different representations of human behavior and also offer different kinds and amounts of evidence for the validity of the tool and technique. There are numerous cases where predictive models have been constructed and used to generate performance predictions, followed by collection of real human performance data for the model context and evaluation of the correspondence between the predicted and empirical data.

But there are very few cases where multiple modeling frameworks have been applied to a complex system application and comparatively evaluated relative to empirical ground truth (with the Air Force AMBR program (Gluck and Pew 2001) representing the principal such endeavor). To our knowledge, there are no cases where such comparative HPM evaluation has addressed the specific requirements of SBA. At the same time, we recognize that system performance data are frequently collected in the test and evaluation (T&E) stage of acquisition, which occurs much later in the development process and typically involves an entirely different team than that which conducted the simulation-based evaluation earlier in acquisition. Because the necessary connections between simulation-based evaluations and empirical T&E evaluations are not always forged at the outset of acquisition, it should not be surprising that comparisons of these two sources of system evaluation data are not typically performed or easily accomplished.

The idea of using HPMs in SBA has actually been with us for a long time, at least since the early 1960s when the Siegel-Wolf task network models were applied to evaluate designs of several major military systems (Siegel and Wolf 1967). Whereas the Siegel-Wolf models were developed by employing a general conceptual methodology in order to implement each model in its own unique code for simulation software (Fortran at the time), subsequent programs in the Navy (with the Human Operator Simulator, HOS; Lane, Strieb, Glenn and Wherry 1981) and in the Air Force (with the Systems Analysis of Integrated Networks of Tasks, SAINT; Chubb 1981) sought to develop general-purpose software tools and environments in order to simplify, aid, and standardize the human performance modeling activities for support of SBA. However, the SBA experience in use of HOS and SAINT, as with the earlier Siegel-Wolf models, was that the costs of using these models was high (with one or more person-year of effort required for typical model applications) and the predictive accuracies were indeterminate and possibly erratic. At least part of the problem was that various other more basic HPMs were typically embedded within the broad scoped applications of tools like HOS and SAINT, each of which introduces additional potential for inaccuracy and unreliability.

There has also long been recognition in SBA communities that it might be appropriate to establish collections or families of complementary HPMs to be used for different aspects and issues of SBA. Such collections were developed by the Navy in its CAFES program and by the Air Force with its CADET program, both in the 1970s, with some components addressing task performance (like HOS and SAINT), others anthropometry (like the Air Force COMBIMAN and the Navy CAR), etc. In the militarily important task domain of manual tracking of dynamic targets (fundamental for many aspects of aircraft piloting and air defense weapons operations), a fairly elaborate genre of mathematical models has been developed in order to predict human performance characteristics from detailed design features of control systems along with environmental dynamics (e.g., McRuer and Krendel 1974). There has also been a long tradition in the development of models for factors that seem to have diffuse moderating effects on virtually all human performance, factors such as fatigue and circadian rhythms, stress, chemical and biological agent effects, performance enhancing/degrading drugs, etc. (e.g., Neville et al., 2000).

Over the past decade, the Army has sought to impose some order on the otherwise seemingly arbitrary processes of selecting and applying human performance modeling tools in support of SBA. In their IMPRINT program (Allender et al., 1995), they have commissioned the development of several simulation-based tools to address different distinct aspects of human accommodation and performance issues pertinent to SBA decisions, including components to address issues associated with personnel selection, training, survivability, workload, and system manning. Use of these tools has been required in the course of recent Army weapons system acquisition contracts. The Air Force has recently implemented its own variant of the IMPRINT task performance modeling tool for its CART program (Brett, Doyle and Hale 2003). This tool employs a task network representation, directly descended from the SAINT tool, to generate dynamically executable descriptions of human performance. This task modeling tool, like all other task network simulations of human performance back to Siegel-Wolf, also offers a library of micro-model and moderator functions to support the construction of human performance simulations. Without any micro-models, the analyst building the simulation would have to specify all of the relevant performance characteristics of each task (e.g., time duration, prerequisites, post-task branching, workload characteristics, etc.). Micro-models provide parametric descriptions of some or all of these parameters. For example, a target detection micro-model might establish a deterministic or probabilistic prediction of the time required for target detection according to scene characteristics and human operator state characteristics. Moderator functions, such as for fatigue or stress, might establish proportional adjustments to all task times, or may possibly make differential adjusts to different tasks or different micro-models within tasks. Another noteworthy characteristic of the IMPRINT task modeling tool is that it offers a substantial library of generic simulation templates for a broad range of typical military weapons systems, such as for a fighter aircraft pilot, tank commander, or anti-aircraft gunner, thus facilitating construction of new applications by starting from a generic template that is fairly close to the concept for the SBA system of interest.

In its recent Agent-based Modeling and Behavioral Representation (AMBR) program (Gluck and Pew 2001), the Air Force has also investigated the relative effectiveness of several competing cognitive modeling architectures for the simulation of detailed human performance characteristics in the context of an abstracted air traffic control task that is similar to many military task environments. As distinct from the task network representations of IMPRINT and CART, AMBR has investigated only a small set of alternative knowledge-based models of cognitive performance involving complex architectures that serve to integrate component functions of attention, memory, perception, decision-making, motor action, and so on. AMBR has investigated the relative effectiveness of four distinct cognitive modeling architectures (ACT-R, SOAR/EPIC, DCOG, and iGEN™) as used by the teams who developed each architecture for predicting a broad range of performance characteristics in the chosen task environment, including the fine details of task dynamics, learning of complex concepts, and transfer of learning to new conditions.

The issue of model validation has been addressed throughout all of these HPM developments from the very beginning. But curiously, there is little that can be said definitively about the validity of any of the complex task performance models (i.e., the task network models and the cognitive architectures), even within the constraint of the specific SBA efforts in which they have been seriously applied. As noted by Young (2003), different kinds of HPMs present different issues and options for validation. And in addition to the various types of HPM, it is also appropriate to recognize that validation efforts must probably be tailored to a defined range of application environments and usages (with major differences across SBA, training system, and decision support system applications) as discussed recently by Campbell and Bolton (in press) under the concept of "application validity." In some cases, human performance data has been collected on a working implementation of the system being modeled and then compared to model data, resulting in judgments being made regarding the closeness of the correspondence. For simple task models such as manual control models (e.g., McRuer and Krendel 1974) and for anthropometric models (e.g., Harris, Bennett and Stokes 1982), these types of validations have been extremely productive and conclusive. The problem for models of complex task performance is that we do not have clear criteria for acceptable prediction accuracy for any of the many aspects of performance that these models might predict (e.g., performance timelines, decisions at choice-points, corollary behaviors such as eye movements, affective characteristics such as workload and stress, etc.). Furthermore, because of the complex character of these models, it is generally impossible to attribute any part of the correspondence between model and empirical data to any particular component, layer, or other aspect of the model application. Thus, it is difficult to infer how to attribute the results of any validation effort to the suitability of the task network or cognitive modeling architecture, or to the skills of the modeling team that constructs each application, or the particular collection of micro-models and moderator function models that are employed for that application, or to the technique employed for the estimation of parameter values for all of the many free parameters incorporated in the model application. Also notably absent from most validation efforts is any systematic consideration of the realistic available alternatives for making the SBA decisions of concern, occasionally considering alternative models as in the AMBR program, but seldom identifying and evaluating the non-model-based analytic techniques that the system analyst might used to inform the same design evaluation decisions.

2 OBJECTIVES

Our fundamental goal is to support and enhance the system performance predictions/evaluations depending in some way on human behavior or performance that must be conducted in the course of deciding between system design alternatives in a succession of levels of detail. We need to know that each model-based prediction/evaluation is good enough so that we make the right design decision with appropriate confidence. But we also want to know that we are not expending any more effort (and other costs) than necessary to insure that the right decision is supported at each stage.

The central objective of the present methodology is to support validations and calibrations of models of complex task performance for SBA so that the results of each validation effort can be used incrementally to inform subsequent decisions about what modeling tools and techniques are most appropriate for each new SBA activity. We will investigate how to attribute validation results to all of the distinct facets of SBA modeling efforts so that for subsequent SBA problems we can better determine what technique should be used, what kinds of analyst skills are needed, what component model elements should be incorporated, and how parameters should be estimated. Clearly, this also entails the development of new guidelines for the collection and analysis of performance data from both models and empirical activities, and the construction of scenarios that will sufficiently exercise model and human participants so that the scope of performance data comparisons will warrant the broad extrapolations of validity that are necessary in order to justify a robust and productive SBA environment.

3 HP-MVP METHODOLOGY

Development of this methodology necessarily begins with an investigation of SBA requirements that are intended to be addressed by HPM tools. Following that, we survey and analyze the various HPM tools and techniques that are available to support these types of SBA needs. Next, we investigate the kinds of data collection that are likely to be feasible in conjunction with military system T&E activities. Finally, we review techniques for comparison of simulation-generated data with empirical T&E data in order to adjust and calibrate the simulation models and to develop conclusions and diagnostic inferences regarding validation. The overall concept of the proposed methodology is illustrated in flow-chart form in Figure 1 as integrated into a skeletal system acquisition process with SBA support. The boxes in the figure that comprise the primary developments of the current efforts are shaded in light blue and are further discussed in the following paragraphs.

3.1 SBA Requirements

The latest revisions to the Department of Defense policy (e.g., DOD 5000.1, 5000.2) make clear the requirement to develop and employ a robust and effective SBA approach to system acquisition, though different approaches are being pursued by each of the services (Von Holle 2004). The role of Modeling and Simulation (M&S) in support of this approach is further specified in DOD 5000.59 and 5000.61. This DoD policy further requires each service component to develop M&S policies and procedures that are consistent with their service-dependent needs. For example, the Air Force Policy Document AFPD-16-10 tailors the DoD M&S instructions to its needs. More directly relevant to the objectives of our effort, the Air Force further defines Validation, Verification and Accreditation (VV&A) policy and procedures in AFI-16-1001.

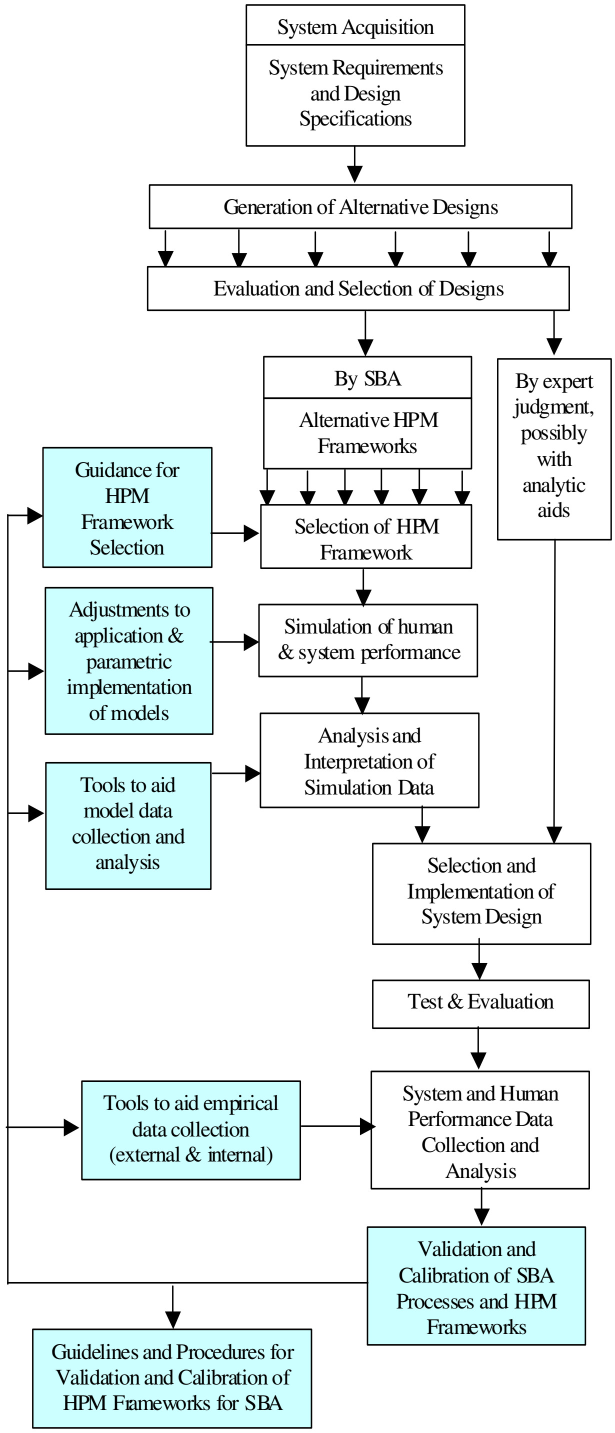

Figure 1: Overview Concept for HP-MVP Methodology

In support of and consistent with these general policies, we are naturally interested in providing a methodology and tools to support a broad range of likely upcoming SBA efforts in association with major new military system acquisitions, not just one or two immediate cases. The diversity across different military systems is considerable in many regards, including especially the kinds of human performance issues that are relevant for critical SBA decisions. In some cases, such as for fighter cockpits, quick visual and manual access to displays and controls can be very important and difficult to achieve, whereas other crewstations pose few concerns in that area. Similarly, some systems present highly stressed workload conditions, whereas others do not. It is important for us to identify a comprehensive range of the kinds of human performance issues that are of concern for SBA decisions across the full variety of military systems. Based on the enumeration of the kinds of military systems to address, we must identify what aspects of performance we want to assess (e.g., performance times, decision quality, workload, situation awareness, confidence, etc.), what performance-influencing factors we want to be able to address (e.g., stress, fatigue, vibration, noise, etc.), and what kinds of individual differences in performance characteristics we need to be able to represent for what populations (e.g., expertise, intelligence, personality, culture, etc.). We must also address the accuracy of prediction that is desired to support the SBA decision, for example by determining how sensitive the SBA decisions are likely to be to variations in predicted performance times or workload scores.

3.2 Survey of Human Performance Modeling Tools and Methods

As noted above, there are many different tools and methods that have been developed to simulate human performance and generate the kinds of predictive data needed for SBA decisions. We have mentioned the task network representations used in IMPRINT and CART, and the knowledge-based representations used in cognitive architectures such as ACT-R, SOAR, EPIC, and iGEN™. Several recent survey efforts have identified and documented the salient characteristics of methods and models applicable to SBA (e.g., Stytz and Banks 2003a, 2003b; Banks and Stytz 2003; Pew and Mavor 1998), greatly facilitating our task to identify these candidates for the present methodology. But we must go beyond this simple identification to determine how these techniques are typically used and what are the components and issues for validation. Do they employ libraries of micro-models or performance moderator functions? Do they offer any guidance, loose or structured, for development of model applications? Are they usable by typical human factors analysts with modest amounts of special training, or do they require extensive special training to the point where they are only used by the tool developers?

In addition to identifying and evaluating the principal human performance modeling methods and tools, it is also necessary to identify some alternative analytic techniques that do not employ human performance models but still address the same kinds of design evaluation questions for military systems acquisition. After all, SBA is a fairly new concept and all of the military services have acquired all of their systems through structured acquisition processes for many decades. Although crude, unaided judgment and intuition have certainly been used in many cases, a variety of ad hoc techniques have been developed for other cases, and a few general analytic techniques to refine human judgment have also been offered. One fairly sophisticated technique for using "anchoring and adjusting" processes to extrapolate from legacy system performance characteristics to proposed new system performance characteristics was offered by the Navy's HARDMAN methodology, which was itself subjected to a fairly elaborate validation study (Zimmerman et al. 1984). Another more recent and more tractable technique was developed in the course of the Navy's Advanced Technology Crew Station (ATCS) program in the form of a Performance Metrics Methodology (Warner, Forster, Messick and Wolf, 1995) that employs a variant of the Quality Function Deployment (QFD) management science technique to generate analytic prediction estimates for the impact of new system design features on human-system performance. But probably the most widely used non-HPM technique for SBA is to use virtual simulations with humans in the loop to obtain empirical data about performance and usability with a new system design. It is important for us to include such alternative techniques with no recourse to human modeling in model validation studies so that we can identify some baseline of prediction performance to which we can compare model predictions. Because the non-HPM-based techniques will generally be much quicker and cheaper to implement than the model-based techniques, it is important to identify domains and requirements for which human-in-the-loop T&E is not feasible, and to determine under what circumstances model predictions are significantly superior to non-model predictions. It is conceivable that we could sometimes find that the models do an adequate job of predicting human performance and supporting SBA decisions, but that unaided expert judgment or expert judgment aided by a simple non-model tool might do just as good of a job at a much lower cost. Validation in this sense must take into consideration all plausible available alternatives for doing the same job.

3.3 Types of Performance Data That Can Be Generated by HPMs and T&E Activities

Human performance can be described in many different aspects and levels of detail according to the behavior representation tool or method being employed. Task network representations can be as simple as identifying just the start and stop times of all of the discrete tasks that make up behavior, but they can also provide a variety of amplifying data and structure, such as workloads and outcomes associated with tasks. Cognitive modeling architectures will typically provide full descriptions of task dynamics along with detailed descriptions of the behavior of component cognitive and manual processes, such as visual perceptions and eye movements. It is appropriate to construct a sort of "common denominator" type of behavioral description language to encompass all of these modeling techniques, such as Ianni (1999) and Badler et al. (2002) have offered for the slightly different context of behavioral representation for graphic human models. This type of common language will be one of the products of this effort and will aid in assessing unique HPMs within the validation framework. Further, the use of this common language to describe model elements will facilitate the establishment of links between T&E HPM requirements and available HPM options. It additionally creates potential for generalizing empirical results across HPMs and HPM elements of the same type, thereby enhancing the efficiency with which HPMs may be validated.

3.4 Technique for Analyzing the Correspondence of HPM and T&E Data

Model validation is much more than the simple comparison of model predictions with empirical data and the binary determination that the model is or is not valid. It is appropriate to view validation on a continuum of processes. At one end of the continuum is model calibration where we use the discrepancies between actual model predictions and empirical data to adjust parametric or structural aspects of the model in order to improve the correspondence for a subsequent execution of the same model. In support of calibration, diagnostic functions are necessary to determine what levels or aspects of the model are responsible for any observed discrepancies. Attribution of distinct aspects of the prediction-observation discrepancies to distinct model facets can serve to facilitate both the calibration of the appropriate facet and the judgment of which facets are working adequately without further adjustment. Without the ability to attribute observed prediction-observation discrepancies to specific aspects of the levels and elements of the HPM processes and tools (i.e., model architecture, knowledge elicitation process, model components, parameter estimation, etc.), it would be very difficult to accomplish incremental improvements on any initial model application. Since major model applications can be fairly costly to undertake, it is essential to be able to have effective techniques for making such incremental refinements in a calibration stage of implementation of the model applications.

At the other end of the continuum, we have fundamental inquiry regarding the inherent value of different modeling frameworks, paradigms, and philosophies. While our methodology acknowledges the need to make such critical judgments, the social and temporal dimensions of such inquiry is well beyond the scope of our present effort. However, between this end-point and simple calibration lies the realm of practical model validations in various forms, depending on the degree of generality and scope of the model being validated. At the least general end of the continuum, the validation of specific model configurations in detailed contexts degenerates into the case of simply adjusting model parameters to achieve a valid model "variant" in a given instance of use (i.e., calibration). Moving toward the more general side of the continuum, increasingly broader classes of models are validated in contexts that are correspondingly more general. For example, a general architecture such as SOAR or iGEN™ might be validated for a class of applications such as pilot vehicle navigation tasks.

The HP-MVP methodology is being developed to provide a framework and scaffolding to promote calibration and validation of human performance models along this continuum. We are developing this framework by working systematically from two given points of reference--the model specification and the empirical performance situation--in order to formulate the integrating representational framework that provides a reliable mapping between the endpoints. At the empirical endpoint, the methodology focuses on a taxonomy of observable actions which permit automated data collection (e.g., keystroke actions, voice utterances, body movements) of data that are conceptually relevant to the model evaluation process. At the modeling endpoint, the methodology identifies events that are 'mappable' to the observable actions and also to evaluation criteria.

While many validation studies have been conducted to calibrate and validate human-performance models against relevant empirical data, the complexity of the many factors and variables involved makes it very difficult to develop general interpretations of the results. One major challenge is presented by the kind of 'bundling' that typically occurs in the development of a human performance modeling application. Typically, the same organization and people who designed and produced the modeling tool/framework are also responsible for the engineering application of the test case, thus making it difficult to attribute any observed results to the modeling tool/framework as opposed to the engineering skills of the project team in accomplishing the immediate application. Also, this same team is sometimes responsible for collection of the empirical data to be used for model evaluation as well as for conducting the statistical analysis and evaluation. Alternatives to this situation are often difficult to arrange because the complexity of the models makes it costly for people not familiar with the models to conduct these types of analysis.

4 CONCLUSIONS

We have described an ongoing investigation into a methodology and supporting tools for validating and calibrating human representations in support of simulation-based acquisition. Continued evolution in the practice of SBA, as well as in human behavior and performance representation technology, requires a general and forward-looking approach to method and tools. The lack of documented experiences in successful validation of human performance models in an SBA context, particularly in the systematic use of later-stage empirical test and evaluation data, implies that our results are preliminary and subject to further iterative refinement through their planned employment and evaluation in a suitable real acquisition process. We have also identified a number of fundamental issues that, while they must be sufficiently addressed in making our prospective methodology and tools operational, will also be subject to ongoing research and change.

ACKNOWLEDGMENTS

The reported work was performed under contract number FA8650-04-C-6438 with the Air Force Research Laboratory monitored by 1st Lt Matthew Eaton, who has provided valuable advice and assistance in this project. The authors would also like to acknowledge the important guidance and technical contribution of Dr. Michael Young (also of AFRL). All judgments and opinions are, however, those of the authors and do not represent positions of the U.S. Air Force.

REFERENCES

Allender, L., , T. D. Kelley, L. Salvi, J. Lockett, D.B. Headley, D. Promisel, D. Mitchell, C. Richer, and T. Feng. 1995. Verification, validation, and accreditation of a soldier-system modeling tool. In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting.* 1219-1223. Santa Monica, CA: HFES.

Badler, N., J. Allbeck, L. Zhao, and M. Byun. 2002. Representing and parameterizing agent behaviors. I*n Proceedings of Computer Animation.* 133-143. Geneva, Switzerland: IEEE Computer Society.

Banks, S.B. and M. Stytz. 2003. Progress and prospects for the development of computer-generated actors for military simulation: part 2— reasoning system architectures and human behavior modeling. *Presence*, 12(4): 422-436.

Brett, B.E., J.A. Doyal, and C.R. Hale. 2003. An analysis system relating individual human performance measures to overall mission effectiveness. In *Proceedings of the Twelfth International Symposium on Aviation Psychology*, Wright State University, Dayton OH. Clearance number ASC: 03-0347.

Campbell, G. E. & Bolton, A. E. (in press). HBR validation: Integrating lessons learned from multiple academic disciplines, applied communities and the AMBR project. In *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation.* ed. R.W. Pew and K. Gluck . Lawrence Erlbaum.

Chubb, G.P. 1981. SAINT, A digital simulation language for the study of manned systems. In Manned Systems Design: Methods, Equipment, and Applications. ed. J. Moraal and K.-F. Kraiss. New York: Plenum Press.

Gluck, K., and R. Pew. 2001. Lessons learned and future directions for the AMBR model comparison project. *Proceedings of the Tenth Conference on Computer Generated Forces and Behavioral Representation.* 113-121. Orlando, FL:SISO, Inc.

Harris, R., J. Bennett, and J. Stokes. 1982. Validating CAR: a comparison study of experimentally-derived and computer-generated reach envelopes. In *Proceedings of the Human Factors Society 26th Annual Meeting*, Santa Monica, CA: Human Factors and Ergonomics Society.

Ianni, J. 1999. A specification for human action representation. In *Proceedings of SAE International Conference on Digital Human Modeling for Design and Engineering.* Warrendale, PA: Society of Automotive Engineers (CD-ROM).

Lane, N., M. Strieb, F. Glenn, and R. Wherry. 1981. The human operator simulator: an overview. In *Manned Systems Design: Methods, Equipment, and Applications.* ed. J. Moraal and K.-F. Kraiss. New York: Plenum Press.

McRuer, D. and E. Krendel 1974. Mathematical models of human pilot behavior. AGARDograph 188. London, UK: NATO Advisory Group for Aerospace Research and Development.

Neville, K., N. Takamoto, J. French, S.R. Hursh, and S. G. Schiflett. 2000. The sleepiness-induced lapsing and cognitive slowing (SILCS) model: Predicting fatigue effects on warfighter performance. In *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society*. 3-57-3-60. Santa Monica, CA: Human Factors and Ergonomics Society

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

Siegel, A.I. and J.J. Wolf. 1967. *Man-Machine Simulation Models*. New York: John Wiley & Sons.

Stytz, M.R. and S.B. Banks. 2003a. Progress and prospects for the development of computer-generated actors for military simulation: part 1—introduction and background. *Presence.* 12 (3): 311–325.

Stytz, M.R. and S.B. Banks. 2003b. Progress and prospects for the development of computer-generated actors for military simulation: part 3—the road ahead. *Presence.* 12 (6): 629-643.

VonHolle, J. 2004. The new SNA–revisited. The MSIAC's Journal On-line, 5 (2) `<http://www.msiac.dmso.mil/journal/joe53_1.html>` [accessed July 15, 2004].

Warner, N.W., E. Forster, M. Messick, and J.J. Wolf. 1995. Performance metrics methodology: bridging the gap between subsystem measures of performance (MOPS) and mission measures of effectiveness (MOEs). *Proceedings of the Eighth International Symposium on Aviation Psychology*, Columbus, OH.

Young, M.J. 2003. Human performance model validation: one size does not fit all. In *Proceedings of the Summer Simulation Conference.* San Diego, CA: Society for Modeling and Computer Simulation.

Zimmerman, W., Butler, R., Gray, V., Rosenberg, L., & Risser, D. T. 1984. Evaluation of the HARDMAN comparability methodology for manpower, personnel and training. Report No: NAS 1.26:173733; JPL-PUBL-84-10. Pasadena, CA: Jet Propulsion Laboratory.

AUTHOR BIOGRAPHIES

DR. FLOYD GLENN is Executive Vice President and Chief Scientist of CHI Systems, Inc. where he leads and oversees projects in the areas of human factors engineering, training technology, decision support and expert systems, systems analysis, human operator modeling, and human performance experimentation. Dr. Glenn has a Ph.D. in mathematical and cognitive psychology from the University of Pennsylvania, an Sc.B. in applied mathematics from Brown University, and is a Certified Human Factors Professional. His primary research interests are in the areas of simulation models of human cognitive and physical performance, performance aids for operators of complex systems, advanced human-computer interfaces based on eye monitoring and voice interaction, and intelligent computer-based training systems. His email address is `<fglenn@chisystems.com>`.

KELLY NEVILLE is a Managing Cognitive Engineer for CHI Systems, Inc.. Her areas of interest include expert-novice differences in complex work environments, stressor effects on cognitive performance, human information processing limitations and strategies, team performance and training, and cognitive work analysis. Dr. Neville holds a Ph.D. in Experimental Psychology with an emphasis on Applied Cognitive Psychology from Rice University and has fourteen years of experience in applied cognitive and human factors research and development. Her email address is `<KNeville@chisystems.com>`.

JAMES M. STOKES is a Managing Software Engineer at CHI Systems. He has led and managed numerous advanced technology research efforts involving both analysis and development. His research interests include simulation-based training, human performance modeling, speech enabled cognitive agents, and multi-modal human-computer interaction. He holds an M.A. in Computer Science from Temple University and an M.A. in Anthropology from the University of Pennsylvania. His e-mail address is `<jstokes@chisystems.com>`.

DR. JOAN RYDER is a Managing Cognitive Scientist at CHI Systems. She has led and contributed to numerous research and development efforts over the last twenty years

involving cognitive modeling, training system analysis and development, user interface design, and human factors analysis. Her interests are in the areas of cognitive modeling, cognitive task analysis, instructional agents, and advanced instructional systems. She holds a Ph.D. in experimental psychology from Brandeis University, an A.B. in psychology from Brown University, and is a Certified Human Factors Professional. Her email address is <jryder@chisystems.com>.