Doesburg, W.A. van (2007). Quality assessment of human behavior models. In 11th IASTED International Conference on Artificial Intelligence and Soft Computing, ASC 2007, 29 August 2007 through 31 August 2007, Palma de Mallorca,

QUALITY ASSESSMENT OF HUMAN BEHAVIOR MODELS

Willem A. van Doesburg
Training & Instruction
TNO Defense, safety & security
Kampweg 5, 3769 ZG
Soesterberg, The Netherlands
email: willem.vandoesburg@tno.nl
www.tno.nl

ABSTRACT

Accurate and efficient models of human behavior offer great potential in military and crisis management applications. However, little attention has been given to the manner in which it can be determined if this potential is actually realized. In this study a quality assessment approach that combines the perspectives of application users, scientific users and model developers is proposed. A case study was done to assess the quality of two models. Specifically the reusability and relevancy of these models was assessed. The assessment approach is a framework in which quality assessment of human behavior representations can be further developed and that offers the exchange of quality assessments results. This, in turn, enables models to be composed of (sub)models of which the quality is established beforehand.

KEY WORDS

Quality Assessment, Human Behavior Modeling, Cognitive Modeling, Decision Support, Cognitive Attention, Tactical Picture Compilation

1 Introduction

Accurate and efficient models of human behavior offer great potential in applications in domains as military operations and crisis management [1], [2], [3]. However, little attention has been given to the manner in which it can be determined if this potential is actually realized. The assessment of the quality of a model is rarely practiced. The few reports on this subject that were found were often anecdotal reports that did not focus on a structured or comprehensive quality assessment. A general quality assessment method is needed in order to learn from past efforts, to leverage existing human behavior representation (HBR) models, and to enable comparison of model quality.

The meaning of quality assessment of HBR models is often limited to the validation of model behavior versus human behavior. Even in this limited scope little experience has been documented [4]. Several studies have proposed that such validation can only be done relative to its application in a certain context [5], [6]. However, these revised validation approaches place little emphasis on non-empirical quality assessment and the meaning of quality for model developers. The notion that model quality is only mediated through behavior validity is too narrow. This interpretation of model quality has meaning for behavior scientists mostly and ignores other perspectives. There is a lack of research in other forms of quality assessment.

HBR quality, similar to validity, can only be viewed relative to the intended use. Moreover, it can only be viewed relative to its intended user. For instance, in current validation, HBR's are evaluated on the appropriateness of simulated human behavior generated by the model. This is mostly relevant to the cognitive scientist interested in the mechanisms underlying human behavior. However, when a model is used in a computer application (e.g. a training simulation or a decision support tool), the application end user is mostly interested in how well the model supports him in reaching his goals (e.g. achieving a learning goal, making better decisions, etc.). Furthermore, the model developer has another distinct perspective on HBR quality. The model developer is mostly interested in how well a model can be used in developing HBR based applications. In this study we aim to define a quality assessment strategy that takes all these users of HBR models into account. In this paper we report on a quality assessment approach that offers a structured way to assess HBR model quality. Additionally, we report our first experiences with this approach during the assessment of the HBR model of a naval command centre officer that was used to create a decision support application for tactical picture compilation. Finally, we discuss the implications of our experience for quality assessments and further work.

2 The present study

In this study HBR models for training and decision-support have been investigated. This study is part of a larger effort in which we aim to create a HBR model architecture that enables the modeler to build a HBR model by composing it from various model components. In order for this architecture to be usable quality information for the available components is required. This allows the cognitive modeler to compare various model components and choose the most qualified.

We define quality as the measure of efficiency and effectiveness. In other words, we want to build good models quickly. Our aim during quality assessment is to identify whether a component contributes to the efficiency of model construction, to the perceived quality of the total system by the end user, or both. The goal in developing our HBR model architecture is to improve model construction by affording the modeler to reuse model components that have a proven positive impact on quality. Models can be created faster when existing model components can be added at a relative low cost to the developer. Moreover, this prevents modelers to have to reinvent wheels. The need for information about the impact of reusing components is further underlined by the Ariane 5 incident. The explosion of the Ariane 5 launcher rocket can be argued to be caused by reusing a proven component (i.e. the guidance system software from the Ariane 4) to adverse effect [7]. To avoid such effects information about the context in which a component may be reused is required.

Additionally, during quality assessment we aim to identify if the component contributes to model effectiveness by affording the model developer to apply the component to a broad range of applications. A component, even though it may be easily added (e.g. to a HBR model for training), is of little use if it can only be used effectively in a specific way (e.g. in a single training scenario).

3 Related Work

The idea of component based design of systems is a familiar one in software design [8]. A quality standard for software has been published by the ISO/IEC [9]. An approach to facilitate leveraging individual component quality to increase overall system quality has been proposed by introducing a framework that enables reasoning about quality aspects [10]. Such a system allows a system developer to reason about the quality gain that the choice for a certain component may yield. The framework contains at the moment only one quality aspect (i.e. system throughput). Various overall system, as well as individual component evaluation and testing strategies have been proposed [11], [12]. However, additional work is required to investigate the relation between individual component quality and overall system quality in HBR models in particular.

The idea of component based design has been translated to the domain of intelligent agents [13]. The approach is centered on formal specification of properties of components. The behavior of the overall system can be verified by observing temporal patterns of behavior of the system and compare these with the requirements that were specified in the design phase of the development process [14], [15], [16]. Although this ensures that requirements made during design are met. It only partly supports the various users of agent systems. When trying to construct an agent from available components little guidance is available to choose one component over another. The verification results of the overall system offers no information about the contribution of individual components. Additionally, the overall system is verified under assumptions about the environment in which the system operates. The relation between individual components and (parts of) the environment is similarly left outside the scope of verification. The emphasis of this approach therefore lies in generating new systems by refining or instantiating components from generic templates rather than reuse of fully specified components.

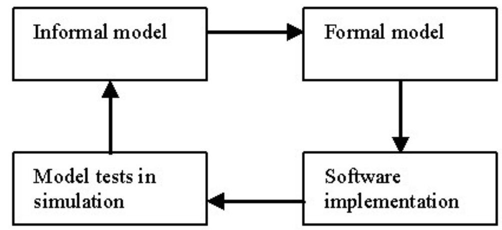

Figure 1. Stages of the HBR modeling process

In the field of psychology the notion of component based modeling is forming around psychological phenomena. A system is proposed in which theories of human behavior variation as a function of certain (internal of external) parameters are expressed as performance moderator functions (PMF) [17]. These functions may be viewed as components that can be used in HBR models to capture specific human characteristics. However, in this system no provision has been made to record how each PMF contributed to the quality of the systems in which it is used and how efficiently it was integrated in those systems.

4 Assessment Challenges

As a discipline in simulation technology, HBR quality assessment is still relatively immature with no theory, few tools and techniques and considerable but poorly documented experience [4]. This is partly due to the unique properties of these models in comparison with other parts of simulation.

- There are multiple perspectives to consider
- They have very high complexity
- There are numerous relationships all interacting chaotically over many different orders of magnitude
- There is a complex coupling between other parts of the simulation system

Part of the complexity arises because quality assessment may be viewed from multiple perspectives. To adequately assess HBR models it is important to have a clear definition of model quality and to have a clear conception how different perspectives relate to model quality. Chandrekarasan &

Josephson [5] have put forward the notion that model quality should be viewed in light of its intended use. This has resulted in a modest following [18], [19]. A model however is not only relevant to end users but also to developers and the researchers. By encompassing these stakeholders in quality assessment process the overall quality is ultimately raised. Model developers can deliver better models faster when they can assess model quality early in the development process. The academic community can supply better theories and methods of human behavior representation if they can adequately assess model performance in relation with human performance. We can therefore distinguish three perspectives on model quality.

- Developer quality (i.e. how effective and efficient is model creation)
- Representation quality (i.e. how well does it represent human behavior)
- Use quality (i.e. how well does it support end user goals)

Various aspects of quality can be distinguished for each of these perspectives. Per aspect there may be one or more associated measures that quantify the quality of the model. To adequately treat the quality of a HBR model from each perspective a little background is needed about the modeling process, see fig. 1.

The HBR modeling process may be divided in four stages, e.g. the modelers in that participated in the AMBR project use this strategy or similar strategies to construct their models for the AMBR project [20]. In the first stage the expert knowledge is elicited from domain experts and organized in a collection of knowledge, called and informal model (usually a collection of interview results).

In the second stage, the modeler extracts the knowledge and strategies from this collection of knowledge and organizes these in a formal model of the expert (e.g. in a bayesian net or a Beliefs Desires Intentions logic). In the third stage the formal description is translated in a computer executable format that can be connected to a simulation of the task environment (or in some cases the real task environment). In the final stage the model is evaluated in the context of the task environment. In the next cycle the issues that may be identified during evaluation are addressed. This cyclical process repeats until the model performs adequately.

Quality assessment may happen in each of these stages. The quality aspects that may be assessed in each stage vary. We can distinguish two classes of quality measures that may be used to measure a quality aspect. These measures can be either empirical when they are used to measure in vivo performance properties of a model. Alternatively, some measures can be conceptual when used to analyze properties of a cognitive model that may not yet be executed by a computer and therefore still resides in a conceptual stage. For an overview of aspects and measures for the three different perspectives on model quality that we propose, see table 1. In general there is no limit on quality aspects that one may assess. Certainly the proposed quality assessment could be extended with more quality aspects, measures and methods. There is a general lack of data concerning model quality assessment and more quality assessments methods and tools are needed.

5 Quality Assessment Approach

We propose a quality assessment approach that takes the challenges described in the previous section into account. In this section we describe the approach. We will focus on the perspective of developer quality. The approach is focused on two aspects of developer quality. First, we focus on the relevance of a model to both the domain it will be used, e.g. commanding a tank battalion or performing a rescue operation, and the application for which it is used, e.g. a tutoring agent, an computer generated (opposing) force. Second, we focus on the reusability of a model, i.e. how well the model can be used to build a novel application.

Relevance is taken to mean the breadth of a model. A model is as broadly applicable as its constraints allow. However, it is difficult to identify the contribution a model provides to the goals of the total system. In literature there is very little data to be found about relevance of HBR models. Quality assessment is mainly an activity performed on cognitive architectures rather than models [1]. Within such assessments the measure of quality is usually defined as the amount correspondence between human behavior and model behavior. Issues such as how general a model is or at what level of detail a model operates are usually glossed over. At the architectural level little can be said about the range of uses a model serves. After all an architecture is only a framework without content and the relevance of the architecture is dependent on how it is leveraged in an individual model.

Relevance, see table 1, is a measure of developer quality. A relevancy score can be assigned to a model. This score is the linear combination of relevancy to a domain task, relevancy to an application and the relative importance of both the domain task and the application to the end user. Two approaches can be identified to construct a model that is relevant for multiple applications in multiple domains. Harmon [21] gives an overview of purposes to which HBR models are relevant in specific simulations. Harmon describes a complete taxonomy of human behavior types that need to be simulated for all tasks and roles of a HBR. Another approach is to incrementally develop models of new behavior types that can be combined and composed with other models to generate behavior. This is the aim of our research.

Domain experts are the source used to determine the relevancy of model. First, we establish the most important tasks that the end user wants to perform and the most important application that the exert uses to perform these tasks. This is investigated by asking the following ques-

Table 1. Overview of quality perspective, quality aspects, measures and assessment methods

Quality type	Quality aspect	Measure	Method type
Developer quality	Relevance	Importance of component for intended use	conceptual
	Reusable	Component dependencies	conceptual
		Environment dependencies	conceptual
	Understandable	Cognitive complexity	empirical
		Native intelligence	conceptual
	Computationally affordable	Resource use	empirical
		Timeliness	empirical
Representation quality	External validity	Goodness of fit	empirical
		Error analysis	empirical
	Face validity	Role-specific measures	empirical
Use-quality	Application validity	Use-specific measures	empirical
	Face validity	Goal-specific measures	empirical

Table 2. The format of the relevancy scoring table

	HBR application X		HBR application Y		Avg.
Domain	Variant 1	Variant..	Variant 1	...	
Task A					Domain Score
Task B					Domain Score
...					...
Avg.	App. Score	App. Score	App Score	...	

tions of the domain experts; A) what is your most important task?, and B) What HBR application will help you most?. The tasks are scored on a five point scale. We call this score the Domain Score. The score for the application is called the Application Score, see table 2. With this information we can determine a relevancy score for a model. This is calculated by taking the average importance of all the tasks and applications for which the model is relevant. After interviewing the domain experts we are left with a table full of relevancy scores. Next, the scores that apply to the model we are assessing are identified by interviewing the HBR model expert. The HBR model expert is asked for which HBR application and which tasks in the domain his model is developed. All scores for those tasks and uses for which the model was not developed do not contribute to it's relevancy score. The final Application Score is determined by taking the average score of those *applications* for which the model was developed. The final Domain Score is determined by taking the average score for a of those *tasks* for which the model was developed. The two relevancy scores may be used to quantify the relevance aspect for a HBR model. These scores offer a means for developers and users of these models to compare different models. However, this requires that the scores are recalculated with every addition or change to the set of tasks and applications.

The other quality aspect we will consider is the reusability aspect. Reuse is an aspect of HBR model quality that relates to its design. Furthermore, reuse is an aspect that lends itself well to conceptual analysis [22]. A good design ensures that model functionality is encapsulated in models. That means that there is an explicit distinction between the description of what a model can do, i.e. its interface, and how it does it, i.e. its implementation. Model capabilities should be properties of the model itself and should have as few dependencies to other models as possible. Dependencies therefore can be used to measure encapsulation of model functionality.

The measure of dependency between system components or models is called coupling [23]. Coupling can be *low* or *high*. Low coupling means that one model does not have to be concerned with the internal implementation of another model, and interacts with another model with a stable interface. Through low coupling, a change in one model will not require a change in the implementation of another model. Low coupling is a sign of a well structured system. However, in order to construct interesting complex systems, decoupling should not be the only design goal: coupling is necessary because interconnected models have more properties and operations than their individual, unconnected elements, leading to a system with (potentially) more functionalities and better performance.

The types of coupling that are relevant in a reusability assessment are, in order of lowest to highest coupling:

- **Data coupling:** Data coupling is when models share data through, for example, parameters. Each datum is an elementary piece, and these are the only data which are shared (e.g. passing an integer to a function which computes a square root).

- **Stamp coupling (Data-structured coupling):** Stamp coupling is when models share a composite data structure, each model not knowing which part of the data structure will be used by the other (e.g. passing a student record to a function which calculates the student's final grade).

- **Control coupling:** Control coupling is when one model controlling the logic of another, by passing it information on what to do (e.g. passing a what-to-do flag).

- **External coupling:** External coupling occurs when two model share an externally imposed data format, communication protocol, or device interface.

- **Common coupling:** Common coupling is when two models share the same global data (e.g. a global variable).

- **Content coupling:** Content coupling occurs when one model modifies or relies on the internal workings of another model (e.g. accessing local data of another model).

The amount of coupling can be quantified by analyzing the HBR application and taking the sum of all couplings. The lower the amount of coupling is, the higher the design quality of the system.

6 Human Behavior Representation Case

In this study quality assessment was applied to a case. The case was used as a means to operationalize our quality assessment strategy and to evaluate the system described in the case. The case concerns a decision support tool that was developed to help officers of the Royal Dutch Navy in their tactical picture compilation duties.

The warfare officer of a frigate of the Royal Netherlands Navy is given the task to find, identify and engage enemy ships. The officer is provided with location, heading and speed information of all the traffic within radar range. The difficulty stems from the limited nature of the information provided and the volume of information that needs to be processed. This yields a high cognitive workload. In order to identify the intent of a ship, the officer first needs to obtain sensor information on a ship. That means that an officer needs to employ the sensors of his ships or to collaborate with other allies to obtain more information. Furthermore, the officer needs to monitor traffic behavior (direction, speed, course changes) closely over a longer time period and integrate this information over time to develop his situation awareness. Additionally, the officer needs to consider his tactical position when contacts are identified as suspect or hostile (based on identification criteria, called IDCRITS). The officer needs to maintain a tactical advantageous position in which he can optimally employ weapons. Furthermore, he needs to take into account the Rules Of Engagement (ROE). These rules specify precisely which (offensive) actions are allowed in each situation. The officers consult these rules to decide when to act.

A cognitive model of an officer that describes how tactical picture compilation is executed can give support on each of these subtasks of the officer. The main bottleneck in the task above is the amount of information that needs to be processed by the single officer. This amount may be reduced by taking over subtasks, or by monitoring them and provide alerts when a task requires attention. In all these cases the cognitive model can only offer sound support when it observes the behavior of the officer as well as the behavior of ships in the surrounding area. The first aspect is implemented by informing the cognitive model about the officer's position of gaze on the tactical display, from which it determines the officer's current task focus. For the latter the model accesses the information on the ships in the area from the tactical display. To support tactical picture compilation it reasons about other relevant knowledge like the ROE, the IDCRITS and sensory systems (i.e. helicopters and maritime patrol aircraft).

The decision support application was implemented (see figure 2) and evaluated on the quality of the tactical picture model and the decision support model.

7 Results

The approach described above was applied to the case described in section six. The HBR application consists of two components.

- A Cognitive Model of Tactical Picture Compilation passing an integer to a function which computes a square root).

- A Cognitive Model of Visual Attention

Both models were assessed on two aspects of developer quality; relevancy and reusability.

The relevancy score was determined by three officers of the Royal Dutch Navy. They were interviewed to determine their most important tasks and to determine what role HBR models can play in these tasks are most important. The tasks that were ranked are based on the Univer-

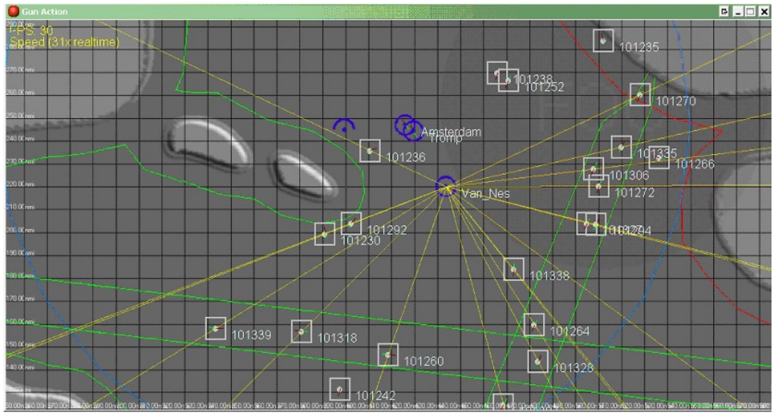

Figure 2. The Tactical Picture Compilation simulation environment

Table 3. Relevancy of models expressed in Application score and Domain score

	Application Score	Domain Score
Tactical Picture Compilation	3.31	1.63
Visual Attention	1.40	3.05

sal Navy Task List for the US navy and correspond with the tasks described in the doctrine of the Dutch Navy. The HBR applications were training and decision-support. The importance values and the relevancy analysis resulted in the application and domain scores for the two models, see table 3. The designers of the models were interviewed to determine the reusability of the two models. During the interview the models were analyzed and the coupling between the models in the system was determined. For the results, see table 4. The dependencies of the models were analyzed in the context of the case at hand. They can and should be analyzed in additional different contexts.

The models were analyzed in relation to each other. The library of components, consisting of two so far, offers only limited evidence. More opportunities to gather reusability evidence will arise as the library expands as well as the number of applications.

8 Discussion

The section above reports a preliminary report on the quality of the HBR model of this study. The investigation of model quality will be an ongoing process in which new model properties and relations between models will be added to the evaluation approach. The scope of the reported assessment is limited to the developer perspective only. Additional quality information needs to be gathered to determine the value of these models for end users and behavior scientists.

The performed evaluation yields evidence in the form of a domain and application score that both model models are of use to the Royal Netherlands Navy. The results reflect the different nature of the models. The visual attention model has a relative low application score and a high domain score. It is relevant for almost all naval tasks but is more suited for decision support applications. The tactical picture compilation model has a relative high application score and a low domain score. This model can be said to be more 'task oriented'. It is relevant for all applications both in training and decision-support but is only relevant for tasks that rely on tactical picture compilation.

During the acquisition of data about model relevancy there was discussion about the tasks for which each model is relevant. The discussion stems from the argument that each model has to be placed in the context in which it will be used. This is called the 'instantiation' of the model. Each model requires different context relevant information in order to function. No clear distinction has been made between models pre-instantiation and post-instantiation. Both models have a different level of abstraction concerning instantiation. It is therefore difficult to compare these models.

Both models seem equally reusable. Investigation of the coupling of models has yielded no constraints on their independent (re)use. However, the decision support tool is still under development. The complete decision support system design will associate the different models with each other. Additionally, a new model may be introduced that combines information about the attention of an officer with task information to judge the appropriate work division between officer and the decision support system. Such a decision making model is likely dependent on both models to

Table 4. Dependencies between tactical picture compilation and visual attention model

Coupling Type	Description
Data coupling	Since the components do not inherently need each other to operate, there is no data passed between components
Stamp coupling	The same holds for data passing through data structures
Control coupling	There is no interdependency on the logic of each component
External coupling	The components are defined independently of each other and have no shared data imposition
Common coupling	The components share no global data
Content coupling	There are no internal interdependencies

feed it with all the necessary information. Furthermore, no models were assessed that concern the reasoning process of the global system. For instance, does the system process information in parallel or serially? What is the strategy that determines the focus of control between models during operations? Such cognitive process models form the basis of the cognition process of the system and are therefore more likely to impose coupling on other models. In general, adding more models increases the likelihood of coupling between models.

This preliminary evaluation of decision support models adds to the existing body of literature on the use of cognitive models in the naval domain. Both HBR models, and the architectures that they are based on, are only rarely investigated on their relevance and reuse properties. Even though there is a large body of literature there is no clear direction in the current community. Harmon [10] shows that a great variety of model categories are relevant to the endeavor of creating HBR models that are successful in providing training capabilities and decision-support. A system is needed to evaluate models and their constituent submodels on their relevance to an application and the opportunities for reuse in order to make the development of these models tractable. Only then can effective modeling collaboration take place and can the community start to leverage existing models for new solutions.

We have made a start in defining a new component based cognitive architecture. As a first step is the development of the two models of this study. This architecture forms the basis of a library of components that can be composed into various applications of cognitive models for naval tasks. There was no pre-existing method to evaluate a such an architecture and its components on quality. This study offers a tentative method to perform such evaluations. This method may be a good departure point to consider accreditation of cognitive models for military purposes. However this requires that the model assessment yields useful information about the quality of cognitive models. Final accreditation could culminate in a judgment about the maturity of a model. This is similar to the Technology Readiness Levels proposed by Meystel, Albus, Messina & Leedom [24].

The models and the evaluation method form the beginning of a comprehensive library of model that can be exploited in various cognitive models. These can subsequently be exploited to perform key roles in training and decision-support. This provides a valuable contribution to the efficiency of training and optimization of operational performance. Clearly indicate advantages, limitations, and possible applications.

9 Further Work

Clearly the models presented here are but the jump-off point for a full fledged agent architecture. However the groundwork has been done and a research framework has been laid down. We will add to our model library by extending our avenue of investigation to other models, such as an instruction model for training. This will additionally require elaboration of the quality assessment method that has been suggested in this report.

An important focus of future effort is to not only develop conceptual model but to test these models empirically by implementing these in functional systems. The tactical picture compilation model and the visual attention model will be combined in a complete cognitive agent for use in a tactical picture compilation decision support system. This enables the decision support concept to be demonstrated and allows the extension of quality assessment to empirical methods.

Acknowledgements

This research was funded by the Royal Netherlands Navy and is part of the Cognitive Modeling research program performed by TNO.

References

[1] Pew, R. W. & Mavor, A. S., *Modeling Human and Organizational Behavior* (Washington DC, National

Academy Press, 1998).
[2] Ritter, F. E., Shadbolt, N. R., Elliman, D., Young, R., Gobet, F., & Baxter, G. D., *Techniques for modeling human and organizational behaviour in synthetic environments: A supplementary review* (Wright-Patterson Air Force Base, OH, Human Systems Information Analysis Center, 2003).
[3] Schurr, N., Marecki, J., Kasinadhuni, N., Tambe, M., Lewis, J. P.,& Scerri, P. The DEFACTO System for Human Omnipresence to Coordinate Agent Teams: The Future of Disaster Response Poster *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, 2005.
[4] Harmon, S. Y., Hoffmann, C. W. D., Gonzalez, Avelino J., Knauf, R. & Barr, V. B., Validation of Human Behavior Representations *Proceedings of Workshop on Foundations for Modeling and Simulation (M&S) Verification and Validation (V&V) In the 21st Century*, Society of Modeling and Simulation (SCS), San Diego, CA, 2002, 1-34.
[5] Chandrasekaran, B., & Josephson, J. R., Cognitive Modeling For Simulation Goals: A Research Strategy for Computer-Generated Forces *Proceedings of the 8th Computer Generated Forces and Behavioural Representation Conference*, 1999, 239-250.
[6] Campbell, G. E., & Bolton A. E., Validating Human Behavior Representations: Moving Beyond P̈reachingT̈o P̈ractice¨*The proceedings of The Interservice Industry Training, Simulation & Education Conference (I/ITSEC)*, Orlando, FL, 2005.
[7] Jzquel, J-M, Meyer, B.: Design by Contract: The Lessons of Ariane, IEEE Computer 30(2):129-130, 1997.
[8] Szyperski, C., *Component software: beyond object-oriented programming* (ACM Press/Addison-Wesley Publishing Co. New York, NY, USA, 1998).
[9] ISO/IEC JTC1/SC7. *Information Technology - Software Product Quality: Quality Model* (ISO/IEC, 9126, 1999).
[10] Sun, C., Empirical reasoning about quality of service of component-based distributed systems *ACM Southeast Regional Conference*, 2004, 341-346.
[11] Balci, O., A methodology for certification of modeling and simulation applications *ACM Transactions on Modeling and Computer Simulation (TOMACS) archive*, 11(4), 2001, 352 - 377.
[12] Briand, L. C., Labiche, Y. & Swka, M. M., Automated, contract-based user testing of commercial-off-the-shelf components *Proceeding of the 28th international conference on Software engineering*, 2006, 92 - 101.
[13] Brazier, F. M. T., Jonker, C. M. & Treur, J., Principles of component-based design of intelligent agents *Data & Knowledge Engineering*, 41(1), 2002, 1-27.
[14] Brazier, F.M.T., Cornelissen, F., Gustavsson, R., Jonker, C.M., Lindeberg, O., Polak, B. & Treur, J. Compositional Design and Verification of a Multi-Agent System for One-to-Many Negotiation *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS'98)*, IEEE Computer Society Press, 1998, 49-56.
[15] Cornelissen, F., Jonker, C.M. & Treur, J., Compositional Verification of Knowledge-based Systems: a Case Study for Diagnostic Reasoning *E. Plaza & R. Benjamins (eds.), Knowledge Acquisition, Modelling and Management, Proc. of the 10th EKAW*, Springer Verlag, 1997, 65-80.
[16] Jonker, C.M. & Treur, J., Compositional Verification of Multi-Agent Systems: a Formal Analysis of Pro-activeness and Reactiveness *W.P. de Roever, H. Langmaack, A. Pnueli (eds.), Proceedings of the International Workshop on Compositionality, COMPOS'97*, Springer Verlag, 1998, 350-380.
[17] Silverman, B. G., Might, R., Dubois, R., Shin, H., Johns, M., & Weaver, R., Toward A Human Behavior Models Anthology for Synthetic Agent Development, *10 th CGF Proceedings*, New York: SISO & IEEE, May, 2001.
[18] Rolek, E., Hughes, T. & Martin, E., A Strategy for Defining Human Representation Requirements *# 9th-CGF-068, Proceedings 9th Conf. on Computer Generated Forces & Behavior Representation*, Orlando, FL, 16-18 May, 2000.
[19] Allsopp, D. J., Capturing the UK MOD CGF Requirement, *# 02-CGF-014 Proceedings od the 11th Conf. on Computer Generated Forces & Behavior Representation*, Orlando, FL, 7-9 May, 2002.
[20] Gluck, K. A. & Pew, R. W., *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation* (Mahwah, NJ: Lawrence Erlbaum Associates, 2005).
[21] Harmon, S. Y., A Taxonomy of Human Behavior Representation Requirements *The proceedings of the 11th Conf. on Computer Generated Forces & Behavior. Representation*, Orlando, FL, 7-9 May, 2002.
[22] Ambrosio, A. & Ruggia, R., Applying Reuse to Conceptual Modeling: An Analysis of Techniques *Invited paper in AFCET* , Paris, July, 1994.
[23] Bruyninckx, H., *Decoupling in complex software systems* , URL: http://www.orocos.org/documentation/decoupling.html, 2003.
[24] Meystel, A., Albus, J., Messina, E. & Leedom, D., Performance Measures for Intelligent Systems: Measures of Technology Readiness *Proceedings of the 2003 Performance Metrics for Intelligent Systems (PerMIS) Workshop*, Gaithersburg, MD., August 16 - 18, 2003.