Validating Virtual Reality Training Simulations

Brad Cain and Lochlan Magee
Defence Research and Development Canada Toronto
Human System Integration Section/Learning and Training Group
1133 Sheppard Avenue West,
Toronto, Ontario, M3K 2C9
CANADA
Brad.Cain@drdc-rddc.gc.ca Lochlan.Magee@drdc-rddc.gc.ca

ABSTRACT

Virtual reality and other scenario oriented computer based approaches are becoming ubiquitous in modern armed forces training to address the gaps between increasing demand of operational tempos and the limitations of traditional training methods. The fidelity of the physical and visual effects in virtual environments has increased dramatically due largely to advances made in the gaming industry that have made such simulations entertaining. Now there is increasing demand for greater fidelity in the representation of computer generated agents within these simulations. Yet, how do we know whether they have any training value? Perhaps a better question is "What level of representation is required of computer based agents to result in effective training?" Unfortunately, most training simulations are assumed to be effective due to compelling visual effects and the belief in "naïve realism"; few have been subjected to rigorous validation and a similar effect may be occurring in the demand for more representative human models. This paper will review some of the validation methods trainers can employ to assess the training effectiveness of simulation based training in general and begin to address how we can determine a suitable investment in human representation to achieve positive training transfer.

1.0 INTRODUCTION

The process of validating Human Behaviour Representations (HBRs) can be, and should be, much the same as validating other types of simulators and simulations, although the complex nature of human models makes them a greater challenge than is typically (now) the case with physics-based models. HBR and typical Virtual Reality (VR) simulations share a common difficulty that make them more challenging to validate than classical simulators, particularly high-fidelity simulators, because they both lack much of the infrastructure that creates the face validity or ecological validity that so often drives the design of the model. In other words, many aspects of the referent for the simulation are deliberately ignored in the simulation model as they are not particularly relevant to the intended application and leaving these aspects out of the model simplifies development to the point where it is practicable or cost-effective (Magee, 2011; Roscoe, 1991). Nevertheless, many of the same concepts that comprise validation of conventional simulators will also apply to validation of VR simulators and the Verification, Validation & Accreditation, Recommended Practice Guide (VV&A, RPG) web page (http://vva.msco.mil/Default.htm) provides many useful, short introductions to the field and a dedicated article on the validation of HBRs (RPG, 2001).

1.1 Validation concepts

In colloquial use, validity has no single, agreed-upon definition but it generally assumed to refer to the degree to which a concept, conclusion or measurement is well-founded and corresponds accurately to the real world. In this sense, it is synonymous with fidelity (RPG, 2000). In the area of scientific research, design and experimentation, validity refers to whether a study is able to scientifically answer the questions it is intended to answer, which has lead to confusion between the terms when used in modelling and simulation.

Unless the modelling objective is to describe mathematically how something works (such as Newton's laws of motion), fidelity is a poor target for M&S validation. Fidelity is ambiguous, although some aspects of fidelity can be quantitatively measured along select dimensions; unfortunately not all of the dimensions of fidelity can be quantitatively assessed. When validation is divorced from the reality label and taken to mean “Fitness for Purpose” much, although not all, of the ambiguity is eliminated. Fitness for Purpose is a human judgement of whether the model or simulation can achieve the desired outcome or accurately predict the occurance of something in a specific application.

Anastasi (1997) describes a number of validation concepts that the human sciences community, in particular psychologists, have been wrestling with for a number of years. Validation in the human sciences is often a more significant scientific challenge than in the physical sciences, because while behaviours are observable, scientists are really trying to understand (model) the unobservable, underlying reasons (latent variables and relationships) that give rise to the boundary conditions (behaviours). Of relevance to HBR, we are trying to validate a sampling of models of human cognition and performance that are based on an incomplete understanding of the referent – the human operators. Thus, validation of HBR models is a somewhat indeterminate science because we are dealing with unobservable, coupled elements with compensatory relationships that often display nonlinear behaviour. In order to tackle this problem, we need to consider a convergence of evidence, continually seeking exceptions that will establish the limits of the applicability of a model. At least three concepts from Anastasi (1997) are relevant to HBR and VR validation. These include the concepts of face validity, content validity and construct validity. Probably more validation concepts should be considered, but often even these three are not dealt with adequately.

1.1.1 Face validity

Face validity assesses the superficial representativeness or appearance of a simulator or simulation, and has been given some fanciful nicknames such as TLAR ("That Looks About Right": Cain et al., 2007) and BOGSAT ("Bunch Of Guys Sitting Around a Table": Campbell & Bolton, 2005). Indeed, some researchers do not consider Face Validity as a scientific concept, yet many simulators developed using finely measured, empirical values matched to governing equations or real-world data subsequently have these values tuned based on Subject Matter Expert (SME) opinion so that the simulation matches their expectations.

Consequently, Face Validity is often the *de facto* limit of validation, even though it may or may not predict effectiveness. In training, face validity can be relevant if it causes the end users to either accept or reject the simulator based on appearance and their belief that they will gain training benefit from it. But Smallman and colleagues (2007; 2005) caution about over reliance on Face Validity as it has a tendency to drive the model or simulator design to unnecessary, or even unproductive, levels of realism (fidelity), often at great cost and expense without a commensurate increase in the Fitness for Purpose. Smallman et al. (2007; 2005) provide examples of where this “naïve realism” (the misplaced belief that increasing fidelity leads to better performance) drive to increase model or simulator fidelity has actually lead to poorer performance than that obtained in simpler, more abstract versions. Similarly, Jacobs (1976) evaluated varying levels of motion base fidelity in aircraft training simulators, where they actually reversed the direction of movement resulting from control inputs, but saw no important transfer to the real aircraft and perhaps more surprisingly, most of the subjects failed to perceived the discrepancy, probably due to the dominance of the visual display channel.

1.1.2 Content Validity

Content Validity is an assessment of whether or not the simulator has all the necessary elements believed to be necessary for the application at hand or according to what is known about the referent. In psychology, Content Validity (also known as logical validity) refers to the extent to which a measure represents all facets of a given construct (a hypothetical relationship among variables, observed or latent). Content Validity can be assessed empirically by comparing the attributes of the resulting device with those in the specification. There is an element of subjectivity that exists in relation to determining content validity, which requires a degree of agreement about the particular traits or characteristics thought to be required, although the judgement is usually based on broad agreement of knowledgeable researchers within the associated field rather than an individual opinion.

Indeed, Content Validity is largely the focus of the High Level Architecture (HLA) verification and validation (V&V) activities: making sure that the specification is sufficiently complete and that the developed product matches that specification (Masys, 2006; RPG, 2001, 2006). This process attempts to establish that “we've built the simulator correctly." But, Content Validity doesn't confirm the simulator's Fitness for Purpose, because Content Validity is usually based on perceived fidelity, i.e., face validity (because of naïve beliefs.)

1.1.3 Construct validity (Fitness for Purpose)

In the psychometric literature, Construct Validity refers to whether a scale measures or correlates with the theorized psychological or scientific construct that it purports to measure. Construct Validity is often confused with “correct implementation of a design” because the word "construct" is being used in a slightly different meaning from its colloquial form. From Wikipedia, the definition for Construct Validity helps to clarify what is meant: "Are we actually measuring (are these means a valid form for measuring) what (the construct) we think we are measuring?"

Usually, Construct Validity is a judgment based on the accumulation of correlations from numerous studies using the instrument being evaluated. This is getting closer to an actual assessment of the Fitness for Purpose of a simulator or simulation. In other words, it attempts to answer the question “Have we built the correct simulator?" and this is where the fun begins…

Determining the Fitness for Purpose of a model, particularly an HBR or VR model, for analytical applications is complicated because of the need to specify a suitable simuland – that is, a specification of the features of the referent (the real object) that the model must represent to a desired level of fidelity. While there are Standard Operating Procedures (SOP) and doctrine that can be used to establish a prescribed behaviour, it is acknowledged that people seldom “follow the rules” and that good decision makers are often guided by SOPs, but are not necessarily bound by them, resulting in unpredictable (or at least unanticipated) behaviours. This means that while SOP rules may be suitable for Computer Generated Forces (CGF) that are intended to act rationally as a collective, they may be inappropriate for representations of the Commanders that direct those groups of CGFs because they don’t capture the individual variability. However, attempting to capture the individual variability complicates the validation process dramatically because the amount of data required for comparison is generally unavailable.

Designing and validating models or simulations intended for analytical purposes is still possible, particularly if a structured design approach is taken, such as decomposing and validating individual elements such as traits and states to moderate behaviour (Content Validation) then comparing the ensemble from a classical statistical approach (Belyavin, 1999; Belyavin & Cain, 2009; Belyavin & Ryder, 2008; Belyavin & Spencer, 2004; Cain & Belyavin, 2008; Cain, Magee & Belyavin, 2011). Decomposing and validating behaviours as done with Hierarchical Task Analysis (Annett, 2003; Annett & Duncan, 1967; Annett, Duncan, Stammers & Gray, 1971; Shepherd, 2000) to produce rational models, provides a foundation that can be tailored to include dependencies upon moderators, producing rational yet plausiblely variable behaviour that can be adapted to the desired objectives.

For training simulations, however, the situation is somewhat different. The degree of goodness or Fitness for Purpose of a simulator, simulation, training method, etc. should be based on the correspondence of the resulting skill or knowledge that is learned to a standard and thus physical fidelity may be less important than psychological fidelity; superficial aspects may be less important than deep construct similarity or process similarity. Training of procedural knowledge and skill-based performance has a ready made metric for validation, Transfer of Training, although a problem does arises as the training objectives themselves become ambiguous requiring scenarios that are less structured and more free-play typical of advanced training.

In general, transfer of training is the effect of previous learning in one environment on performance in another. Positive transfer is the improvement of skill in the operational environment while negative transfer is the development of bad habits that impair performance in the operational environment. Several techniques for assessing training transfer are discussed in an AGARD (1980, Appendix A) publication as well as in several papers by Roscoe and colleagues (Roscoe, 1971; Roscoe & Williges, 1980). Two of the techniques that are particularly relevant for assessing the Fitness for Purpose are Reverse (Backward) Transfer of Training and Forward Transfer of Training.

1.1.3.1 Forward Transfer of Training

A Forward Transfer of Training study is the gold standard for assessing the Fitness for Purpose of a training simulator. The assessment itself has the greatest ecological validity, as the testing occurs in the operational environment (such as a field training exercise), but it carries an associated risk that may result from some unforeseen, negative transfer. Because Forward Transfer of Training studies often require operational equipment in the field, they are expensive to conduct and typically offer limited experimental control that may restrict them to observational studies or studies with small numbers of subjects, so reliability of the results can be a constraint on the generalizability of the observation. A variation on this theme is a Quasi Forward Transfer of Training in which performance is evaluated in a "high fidelity" simulator after training in another environment.

A simple, Forward TOT study comprises two groups: one group that receives training in the simulator being evaluated (Experimental Group) and one group that receives an unrelated treatment over a similar interval (Control Group). Performance in the operational environment is then evaluated, typically following a conventional training plan, until predetermined performance criteria are met, and the differences between the two groups on these performance variables (e.g. number of errors, precision, etc.) are used to assess the effectiveness of the simulator training treatment.

In the Forward Transfer of Training approach, the time spent (or number of trials required) to reach criterion in the operational environment can be the metric used to assess the simulator effectiveness, although other metrics may be used to provide additional insight. The time saved in reaching criterion in the operational environment as a result of the training intervention can be used to calculate a Training Effectiveness Ratio (Roscoe, 1971; Roscoe & Williges, 1980) that is a relative measure of how much field time is saved by using the simulator training treatment. If estimates are available for the hourly use of the field equipment and the simulator, then these data also provide a cost-effectiveness measure of using the simulator in place of the field equipment.

Roman & Brown (2008) cited work by reported by Hill[1] that showed substantial benefits of simulator based training prior to conventional, operational environment training using the pass rate (achieving criterion) at selected points in the program. As presented, there is limited opportunity to conduct statistical evaluation of these data, however, with access to the original data?, the distributions of the various groups can be assessed statistically for significant differences, for example, using nonparametric techniques such as the Mann-Whitney *U* test or parametric tests from the General Linear Model. Estimates of the cost-benefit can be made knowing the hourly cost of the operational equipment and the simulator, then computing the number of operational hours saved by the simulator training by the Experimental Group less their simulator costs.

[1] Hill, L. (2008) Use of Simulation in the Armour School. Military Modelling and Simulation Symposium, SpringSim, Ottawa, Ontario.

1.1.3.2 Reverse Transfer of Training

In any training, whether it is in an operational environment or in a simulated environment, operators may learn inappropriate or undesirable behaviours. This may occur because of the unpredictable nature of the training environment or because the training environment actually elicits specific behaviours unintentionally. Obviously, instructors wish to minimize the opportunity for such negative training transfer to result in harmful consequences such as an accident and ideally, not occur at all. Reverse Transfer of Training is an intermediate activity that provides an estimate of the expected effectiveness of the simulated training environment without incurring the risks associated with training in the operational environment. This is particularly important in military applications.

In a Reverse Transfer of Training study, operators who have demonstrated mastery of the desired skill or knowledge (Experts) perform the training task as do inexperienced operators who have to learn the training task (Novices). The difference in the performance between the Expert group and the Novice group is an indicator of the effectiveness of the simulator.

If a simulator is effective for training a task (i.e. Fit for Purpose), then a certain observations are hypothesized as the two groups repeat the task in the simulator:

1) The Expert group will start at a high level of task performance and will quickly asymptote to the criterion performance level.
2) The Novice group will start at a low level of task performance and then gradually improve over time, converging on the criterion performance level.

If certain unexpected observations occur, then they are assumed to be indicators that a simulator is not effective at training the task (i.e. not Fit for Purpose):

1) The Expert group starts at a low level of task performance (simulator may not provide adequate cues).
2) The Expert group does not improve with practice.
3) The Expert group starts with perfect performance or the Novice group starts with high task performance (the simulator may be too simplistic or the task may not be sufficiently sensitive).
4) The Novice group does not improve with practice.

Because Reverse Transfer of Training is an experiment that can be conducted within a school house or laboratory, it can provide an opportunity to use a greater number of subjects, which affects the statistical reliability of results. Further, because Experts are used as part of the subject pool, there is an inherent opportunity to assess the Face Validity of the training environment and to identify many shortcomings or limitations that could lead to negative transfer.

Reverse Transfer of Training is not a full answer, however, and the actual desired outcome (Positive Forward Transfer) is not guaranteed. While Reverse Transfer of Training should not be the sole justification for claiming that the simulator is Fit for Purpose, it provides an environment that is less risky and allows for more experimental control. Further, it would be unlikely that a simulator would be accepted as a training device if it could not be adequately used by proven experts.

We conducted a small, team training study with the Canadian Forces (CF) Maritime Helicopter (MH) community to assess a proposed training approach for Landing Signals Officers (LSO) that demonstrates the use of Reverse Transfer of Training for validation (Cain et al., 2011). The task selected for the study was predominantly the procedural aspects of the LSO’s task, interacting verbally with other members of the MH

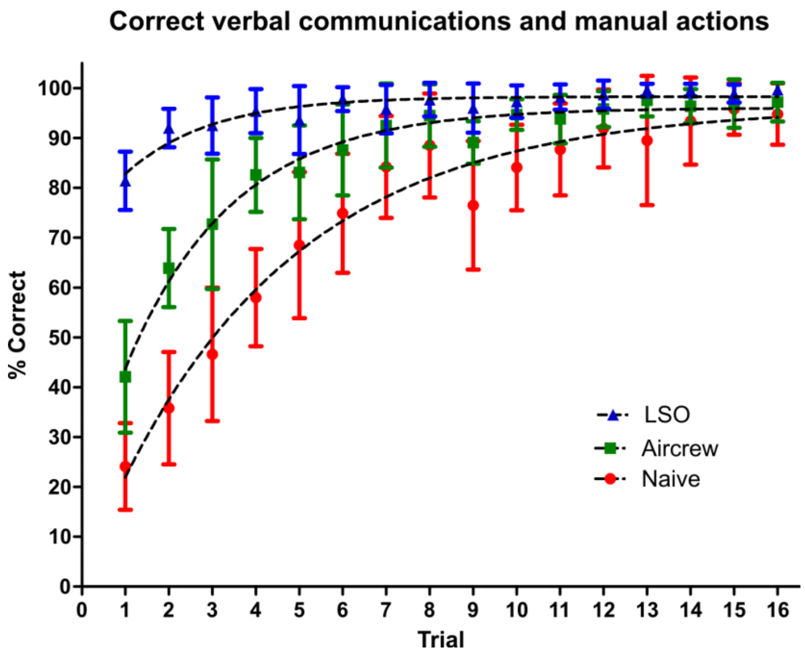

Figure 1. Learning curves three groups of subjects from an LSO training task. Values are group means and standard deviations. n=10 subjects/group. (Cain et al., 2011)

team and the ship's crew, all of whom were HBRs for this study. Three groups of 10 subjects each repeatedly performed the LSO's role during a simulated approach and landing of a CF Sea King helicopter onto a CF Halifax Class frigate under way. There was an Expert group comprising experienced LSOs, a Naive group comprising subjects with no prior exposure to the task, and an intermediate group comprising MH Aircrew who were familiar with the MH environment and had incidental exposure to the task, but who had not been formally trained or performed the LSO's task previously. The subjects had to learn the verbal commands, manipulation of the LSO control console and timings so that they could conduct the procedure without error. Subjects received corrective feedback both during and after each trial.

The principal results from the study (Cain et al., 2011) are shown in Figure 1 where it can be seen that the Expert and Naïve groups perform as hypothesized for a simulator that is Fit for Purpose. Also, as might be expected, the Aircrew group performance falls intermediate to the other two groups, which is interpreted as reflecting the Reverse Transfer of prior knowledge gained from observing landings while at sea.

Several of the Expert group members' LSO qualifications were not current at the time of the study, which lead to a somewhat low initial performance; however, the Expert group quickly reached effectively perfect performance. The data were analyzed as a 3x16 (Group by Trial) mixed Analysis of Variance (ANOVA) model which indicated that the groups were statistically different and that the

difference persisted between the Expert and Naïve groups until about the 12th trial. The data were also fitted to an exponential function and the resulting regression coefficients analyzed, revealing that the groups differed because of their prior knowledge (as expected) but there was only minor (although statistically significant) differences in their learning rates.

2.0 CONCLUSION

In summary, many M&S validation efforts focus on model or simulation fidelity and matching specifications as measures of goodness with the naive assumption that replicating the referent to the greatest extent possible will lead to the best simulator. Yet modelling and simulation is really about abstracting and representing the essential aspects of the referent within a model to meet the objectives of the application. Assessing the Face Validity and the Content Validity are important aspects of any design validation process, but they are not sufficient or always informative Further, it is impossible that complete Content Validity will be established, although "good enough" is certainly feasible.

Fitness for Purpose is a more appropriate and more attainable metric for validation as it relates directly to the effectiveness in the target application rather than attempting to reproduce reality to some arbitrary degree of similarity. This does not mean that simulations, including HBRs, need not be concerned with fidelity, only that fidelity needs to be subservient to fitness for purpose as demonstrated through application.

In some respects, M&S validation for training is conceptually easier than for analytical purposes as there is a ready made metric, Training Transfer, and established methods for evaluating it that do not require arbitrary assumptions about the level of fidelity required. Judgement about the adequacy of the training is still required, however, as resources are limited and so training must proceed at an acceptable rate. Similarly, comparing model performance against observation to assess its Fitness for Purpose in an analytical application should follow traditional scientific methods to determine differences, followed by a judgement about whether detected differences are worth worrying about.

Adopting Forward and Reverse Training Transfer methods for validation provides a structured approach for predicting the effectiveness of a simulator or simulation in the intended application, at least for training simulations. This may not be true for analytical applications, and so some other philosophy will be required to determine "how much" is "good enough." Thus, there are scientific issues that remain for validations in each of these application areas. One such scientific question that is currently the focus of the authors' work is the correlation between Forward and Reverse Training Transfer; that is, establishing the validity of the Reverse Transfer of Training method.

Validation is not a statistical test or a mechanical procedure; validation is a human judgement of adequacy in a given application. As the mathematician George Box (1979) is often quoted, "All models are wrong; some models are useful." Fidelity is a Siren's call that should be muffled by Fitness for Purpose when determining the validity of a model or simulation.

3.0 REFERENCES

[1] AGARD. (1980). *Fidelity of simulation for pilot training* (No. ISBN: 92-835-1 377-0). 7 Rue Ancelle - 92200 Neuilly-Sur-Seine, France: North Atlantic Treaty Organization (NATO), Advisory Group for Aerospace Research and Development (AGARD).

[2] Anastasi, A. (1997). Validity: Basic concepts. In A. Anastasi & S. Urbina (Eds.), *Psychological Testing* (7 ed., pp. 113-139). New Jersey: Prentice Hall.

[3] Annett, J. (2003). Hierarchical task analysis. In E. Hollnagel (Ed.), *Handbook of cognitive task design* (pp. 17-35): Lawrence Erlbaum.

[4] Annett, J. & Duncan, K. D. (1967). Task analysis and training design. *Occupational Psychology, 41*.

[5] Annett, J., Duncan, K. D., Stammers, R. B. & Gray, M. J. (1971). *Task Analysis*. London: Her Majesty's Stationery Office.

[6] Belyavin, A. (1999). *Modeling the effect of stress on human behavior*. Paper presented at the Proceedings of the 8th Conference on Computer Generated Forces and Behavior Representation, Orlando, Florida 481-487.

[7] Belyavin, A. & Cain, B. (2009). *An example of validating models of continuous processes.* Paper presented at the 18th Annual Conference on Behavior Representation In Modeling and Simulation, Sundance Resort, Utah.30 Mar – 2 Apr. 8.

[8] Belyavin, A. & Ryder, J. (2008). *Modeling the effect of changes in retention on performance.* Paper presented at the Behaviour Representation in Modeling and Simulation, Providence, Rhode Island 9.

[9] Belyavin, A. J. & Spencer, M. B. (2004). Modelling performance and alertness: The QinetiQ approach. *Aviation, Space, and Environmental Medicine, 75*(3), A93-A103.

[10] Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201-236). New York: Academic Press.

[11] Cain, B., Armstrong, J., Allender, L., Belyavin, A., Castor, M., Fraser, W., Gluck, K. A., et al. (2007). *Human Behaviour Representation in constructive modelling.* Paper presented at the HFM 143/RSM "Human Behaviour Representation in Constructive Modelling" Canadian Forces College, Toronto, Ontario, Canada.30-31 May. 165.

[12] Cain, B. & Belyavin, A. (2008). *Inclusion of Human Behaviour Representation in military modelling and simulation.* Paper presented at the MSG-060 Symposium on "How is Modelling and Simulation Meeting the Defence Challenges out to 2015?" Vancouver, British Columbia 17.

[13] Cain, B., Magee, L. & Belyavin, A. (2011). *Validating an HBR simulation for training within a virtual environment.* Paper presented at the Behavioral Representation In Modeling and Simulation (BRIMS 2011), Sundance Resort, Utah.22-24 March. 8.

[14] Campbell, G. E. & Bolton, A. E. (2005). HBR Validation: Integrating Lessons Learned From Multiple Academic Disciplines, Applied Communities, and the AMBR Project. In K. A. Gluck & R. W. Pew (Eds.), *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation* (pp. 365-395). Mahwah, New Jersey: Lawrence Erlbaum Associates.

[15] Jacobs, R. S. (1976). *Simulator cockpit motion and the transfer of initial flight training* (No. ARL-76-8/AFOSR-76-4; AFOSR-TR-77-038): Air Force Office of Scientific Research, Air Force Systems Command, United States Air Force.

[16] Magee, L. E. (2011). *Virtual Reality (VR) as a disruptive technology* (Technical Memorandum). Toronto: Defence Research and Development Canada Toronto.

[17] Masys, A. J. (2006). *Verification, Validation and Accreditation*. Paper presented at the IEEE Canadian Conference on Electrical and Computer Engineering. from http://www.ieee.ca/ccece06/

[18] Roman, P. A. & Brown, D. (2008). *Games - Just how serious are they?* Paper presented at the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC).

[19] Roscoe, S. N. (1971). Incremental Transfer Effectiveness. *Human Factors, 13*(6), 561-567.

[20] Roscoe, S. N. (1991). *Simulator qualification: Just as phony as it can be*. Paper presented at the International Journal of Aviation Psychology.

[21] Roscoe, S. N. & Williges, B. H. (1980). Measurement of Transfer of Training. In S. N. Roscoe (Ed.), *Aviation Psychology* (pp. 11): Ames: Iowa State University Press.

[22] RPG. (2000). Fidelity, *RPG Special Topic* (pp. 11): Verification, Validation and Accredition, Recommended Practices Guide.Retrieved 17 Nov 2011 from http://vva.msco.mil/Special_Topics/Fidelity/default.htm

[23] RPG. (2001). Validation of Human Behavior Representations, *RPG Special Topic* (pp. 47): Verification, Validation and Accredition, Recommended Practices Guide.Retrieved 17 Nov 2011 from http://vva.msco.mil/Special_Topics/HBR-validation/default.htm

[24] RPG. (2006). The V&V Agent's role in the VV&A of new simulations, *RPG Core Document* (pp. 68): Verification, Validation and Accredition, Recommended Practices Guide.Retrieved 17 Nov 2011 from http://vva.msco.mil/Role/VVAgentNew/default.htm

[25] Shepherd, A. (2000). HTA as a framework for task analysis. In J. Annett & N. A. Stanton (Eds.), *Task analysis* (pp. 7-24). London: Taylor & Francis.

[26] Smallman, H. S., Cook, M. B., Manes, D. I. & Cohen, M. B. (2007). *Naive realism in terrain appreciation.* Paper presented at the Proceedings of the 51st Annual Meeting of the Human Factors and Ergonomics Society, Baltimore, MD,.Oct 1-5.

[27] Smallman, H. S. & St. John, M. (2005). *Naïve Realism: limits of realism as a display principle.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting, Santa Monica, CA