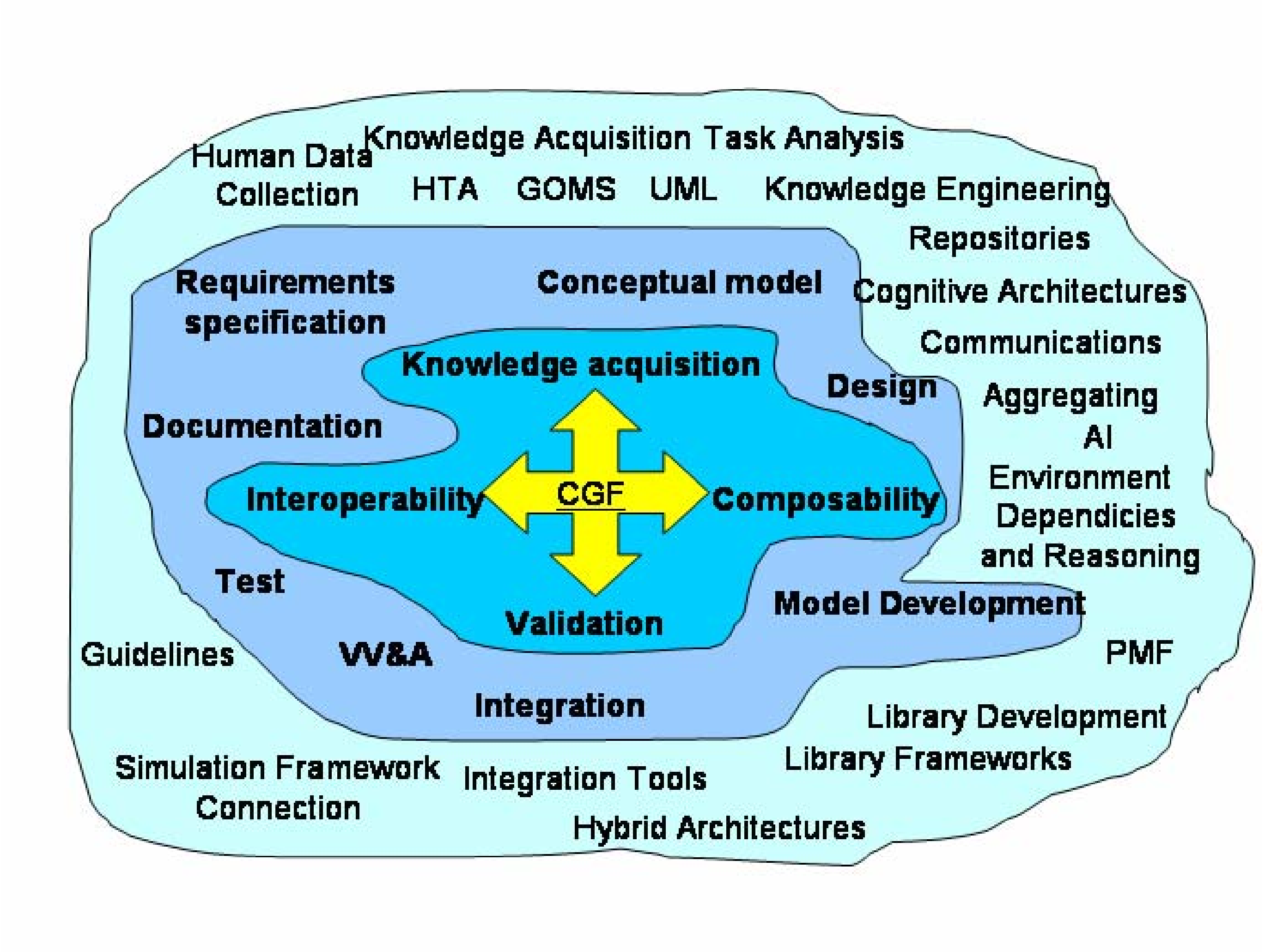
FOI
Defence Research Agency
Weapons and Protection
SE-164 90 Stockholm

Phone: +46 8 555 030 00
Fax: +46 8 555 031 00

www.foi.se

VV&A of CGF: A REVVA Application?

STEN-ÅKE NILSSON, CHOONG-HO YI



FOI-R--2117--SE
ISSN 1650-1942

User report
November 2006

Weapons and Protection

Sten-Åke Nilsson, Choong-ho Yi

VV&A of CGF: A REVVA Application?

FOI-R--2117--SE
ISSN 1650-1942

User report
November 2006

Weapons and Protection

Issuing organization FOI – Swedish Defence Research Agency Weapons and Protection SE-164 90 Stockholm	**Report number, ISRN** FOI-R--2117--SE	**Report type** User report
	Research area code 2. Operational Research, Modelling and Simulation	
	Month year November 2006	**Project no.** E 62144
	Sub area code 21 Modelling and Simulation	
	Sub area code 2	
Author/s (editor/s) Sten-Åke Nilsson Choong-ho Yi	**Project manager** Choong-ho Yi	
	Approved by Nils Olsson	
	Sponsoring agency FMV	
	Scientifically and technically responsible	

Report title

VV&A of CGF: A REVVA Application?

Abstract

Reliability, suitability and validity are some of the most important attributes in CGF (Computer Generated Forces) development and deployment. Because of the inherent complexity of human behaviour, resource and cost driven model development reasons and application diversity these key issues are however so far at a stage of infancy. A growing understanding in the CGF and HBR (Human Behaviour Representation) community is that the efforts to model CGF and human behaviour need tailoring with respect to capabilities and constrains given by the requirements specifications of the application. This further stresses the need to explain suitability and validity aspects. To meet this need efficiently appropriate VV&A methodologies are required.

This report in short describes a recently developed VV&A methodology, the REVVA methodology, gives an overview of current verification and validationen methods and techniques used for the CGF and discusses of the REVVA method for CGF validation. It finally describes some considerations for the road ahead.

To provide an understanding of the problem of CGF validation a review of human behaviour's history from the psychology has been performed, including the efforts to formulate, declaire and model these phenomena. These theories constitute the foundations for cognitive architectures developed during the last three decades. They are the basis for building simulation models of human behaviour today.

This report can not proclaim the REVVA methodology to be the solution for CGF validation. However, useful guidelines are provided by the methodology, e g to build the ToA product and use the tailoring method.

Keywords

VV&A, M&S, CGF, HBR, Validation, Verification, Validity, VV&A Methodology, REVVA

Further bibliographic information	**Language** English
ISSN 1650-1942	**Pages** 29 p.
	Price acc. to pricelist

Utgivare
FOI - Totalförsvarets forskningsinstitut
Vapen och skydd
164 90 Stockholm

Rapportnummer, ISRN
FOI-R--2117--SE

Klassificering
Användarrapport

Forskningsområde
2. Operationsanalys, modellering och simulering

Månad, år
November 2006

Projektnummer
E 62144

Delområde
21 Modellering och simulering

Delområde 2

Författare/redaktör
Sten-Åke Nilsson
Choong-ho Yi

Projektledare
Choong-ho Yi

Godkänd av
Nils Olsson

Uppdragsgivare/kundbeteckning
FMV

Tekniskt och/eller vetenskapligt ansvarig

Rapportens titel

VV&A av CGF: Går REVVA att använda?

Sammanfattning

Vederhäftighet, lämplighet och giltighet är nyckelord för CGF utveckling och användning. På grund av komplexiteten hos utvecklade modeller, höga kostnader för utveckling och ett mycket varierat användningsområde är dock dessa nyckeltermer hittills tämligen förbisedda. Det börjar dock växa ett intresse bland CGF-utvecklare för att något måste göras för att bättre kunna hantera utveckling och användning beroende på kravspecifikationer och användningsområden. Detta sätter också ytterligare krav på att kunna beskriva och mäta ovanstående nyckelord och för detta kunna använda en tillräckligt bra VV&A-metodik som hjälp.

Denna rapport beskriver relativt kortfattat den nya REVVA-metodiken, sammafattar de nu använda metoderna och teknikerna för verifiering och validering vid utveckling av CGFer, resonerar om möjligheten att använda REVVA-metodiken samt visar slutligen förslag på hur arbete inom detta område bör fortsätta.

Rapporten innehåller också ett avsnitt som redovisar hur man inom psykologin under många år arbetat med att utforma teorier och modeller för mänskligt beteende och hur man försökt definiera begreppet giltighet. De framtagna teorierna ligger idag till grund för de ramverk som utvecklats och med vars hjälp man på ett bra och relativt enkelt sätt kan bygga modeller av mänskligt beteende (CGFer).

Denna rapport kan med den information som vi har idag inte entydigt påstå att REVVA-metodiken är till alla delar användbar för validering av CGFer. Delar som t ex ToA och metodiken för att skräddarsy modeller efter behov bör dock vara direkt tillämpbara.

Nyckelord
VV&A, M&S, CGF, HBR, validering, verifiering, giltighet, VV&A metodik, REVVA

Övriga bibliografiska uppgifter

Språk Engelska

ISSN 1650-1942

Antal sidor: 29 s.

Distribution enligt missiv

Pris: Enligt prislista

Contents

1 Introduction ... 2
1.1 The intricate questions ... 2
1.2 Background ... 2
1.3 Purpose and method ... 3
2 The REVVA methodology ... 4
2.1 Basic terminology ... 4
2.2 The “three pillars” ... 5
2.3 Organisations: Parties and roles ... 5
2.4 The REVVA Generic Process (RGP) and the associated products ... 6
3 CGF and HBR ... 10
4 VV&A ... 13
4.1 Verification and Validation of Models and Simulation Results ... 13
4.1.1 Definitions ... 13
4.2 Distinction between validating a model and an architecture ... 14
5 VV&A of HBR ... 15
5.1 The challenge ... 15
5.2 Referents for HBR Validation ... 15
5.2.1 What purpose does a referent for an HBR serve? ... 15
5.2.2 What referents exist for validating HBRs? ... 15
5.2.3 Levels of referents for HBR validation ... 15
5.3 Current practice and deficiencies ... 17
5.4 What is currently available: technology resources ... 17
5.4.1 Tools and techniques ... 17
5.4.2 KBS VVE&T Technology ... 18
5.4.3 Human Behaviour Science ... 19
5.4.4 Useful Techniques for collecting evidence ... 20
5.5 Different types of validity. ... 21
5.5.1 Construct validity and cognitive models ... 21
5.5.2 Application validity: Assessing a model for its intended use ... 21
5.6 V&V activities ... 22
5.7 Emerging Approches ... 22
6 REVVA for CGF ... 24
6.1 Process viewpoint ... 24
6.2 Product viewpoint ... 24
6.3 Organisation viewpoint ... 25
6.4 Evaluation ... 25
7 Conclusions and future work ... 26
Acronymes ... 27
References ... 28

1 Introduction

1.1 The intricate questions

- Given that a human behaviour model is validated for one application what risk do I take if I use that model in another context, with slightly different circumstances, conditions and dependencies?
- When it comes to human behaviour there may be different ways to reach a goal or fulfil a mission. All of the different ways could be fully acceptable, even though only one of them probably is the most efficient one. How can I validate such a model that allows, and should allow, different courses of events? That is, how can I validate Human behaviour?

As pointed out by Gluck and Pew [Gluck & Pew, 05] it is generally agreed that validation is tremendously important, and the risk of drawing erroneous conclusions from unvalidated models are unacceptable. However, validation is difficult, costly and is rarely, if ever, done.

Why is this? Dilemmas facing model developers include selecting from a wealth of validation strategies and a lack of standards by which to judge validation evidence [Harmon et al., 02].

1.2 Background

VV&A (Verification, Validation and Accreditation) is a multi-disciplinary research field aiming at increasing the credibility, i.e. correctness and validity, of simulation models. The increasing use of simulation models and growing complexity of the models put great demands on VV&A. One of the most serious problems in the VV&A field today is the lack of common VV&A methodology. For example, there is no single commonly accepted definition for “verification” and “validation”, two most fundamental terms in VV&A. This implies that people may have different understanding and expectation of VV&A concerning, e.g. the meaning, scope, target and responsibility. Furthermore different organisations quite often use different (“local”) methods and techniques for their VV&A activities. This makes it very difficult for an outsider to understand and judge the credibility of their VV&A findings. Consequently it hinders reuse and exchange of simulation models between different organisations and nations.

Motivated by the situation, several international VV&A groups have been working to develop international VV&A standards. For example, NATO Modeling and Simulation Group, MSG 019 (Activity TG-016 VV&A of Federations), ITOP (International Test Operation Procedure) on V&V [ITOP, 04] and REVVA (THALES JP 11.20 and its follow-on project EUROPA 111-114)[1]. These approaches have different focus. The NATO approach focuses specifically on VV&A for federations being developed according to the FEDEP [FEDEP] where VV&A activities are specified as an “overlay” process to FEDEP [PDG, 06]. The ITOP approach aims at supporting the exchange of V&V information and provides a workbook and structure for documenting V&V information [Sullivan & Chew, 05]. The REVVA methodology is intended to provide a generic VV&A framework which is not restricted to specific M&S (modelling and simulation) development process like FEDEP, or specific application domains, see the next chapter. Sweden has had and still has representatives in all working groups mentioned above, from FMV (Swedish Defence Materiel Administration) and FOI (Swedish Defence Research Agency).

An important aspect to be considered when developing a methodology such as a VV&A standard is the evaluation of the methodology, desirably before it is finalised. For example, by testing it in reality, or comparing it with other methodologies. It is of course true for technique development as

[1] The REVVA will be presented in more detail not only as a project but also as a methodology in the next chapter.

well. By conducting evaluation there is another chance to identify and take care of the drawbacks of the methodology, and improve it. Also, a result may be deeper understanding of it.

1.3 Purpose and method

The purpose of this study, commissioned by FMV, is to evaluate the REVVA methodology which is expected to be finalised by the end of 2008. The evaluation will be made by applying the REVVA methodology to the development and validation of CGF. Methodology for development and employment of CGF or HBR has been studied for more than five years at FOI. This evaluation of the suitability of the REVVA methodology for VV&A of CGF is made with current experience of our work in the CGF development project and from the methodology work in REVVA. The main part of this document is summaries of current knowledge from the appropriate areas.

The expected benefits from this evaluation are twofold. Firstly, from the perspective of REVVA, it is expected to identify the strengths and weakness of the REVVA as a VV&A methodology in general, but also as a VV&A methodology supporting the conceptual modelling in particular. Secondly, from the HBR development perspective, the REVVA methodology could illuminate and contribute to improving the VV&A aspect in the proposed framework.

In addition to HBR development and in parallel with this study, the REVVA methodology has been evaluated on another research project DCMF (Defence Conceptual Modelling Framework) at FOI as well. It was also commissioned by FMV, and is described in a separate report [Yi et al., 06]. These two evaluations with different focus and within different domains provide valuable lessons learned for improvement of the REVVA methodology.

This report is structured as follows. Chapter 2 and Chapter 3 are devoted to presenting REVVA in detail and a summary of CGF (or HBR), respectively. REVVA development is presented as project and also as methodology. Chapter 4 contains some more terminology and definitions of terms that is relevant for the following chapters, for instance a few other definitions of verification and validation are cited and a section about the differences between validating simulation models and specific architectures respectively. Chapter 5 describes the VV&A aspects identified in HBR, i.e. the need, scope and target of VV&A of HBR. The REVVA methodology is then analysed and evaluated in Chapter 6 from the viewpoints of organisation, products and process based on the discussion in chapter 5. Conclusions and future work are provided in Chapter 7.

2 The REVVA methodology

The European project THALES JP 11.20, “Common Validation, Verification and Accreditation Framework for Simulation”, with the working name “REVVA”, has been running since March 2003 through September 2004.[2] The purpose of this project was, as its name implies, to develop the basis for a common methodological framework for VV&A of simulation models and simulation results. REVVA was funded by five nations: France (lead nation, ONERA), Denmark (UNI-C), Italy (DATAMAT), the Netherlands (TNO) and Sweden (FOI and FMV). The REVVA research effort relied on past and existing efforts coming from many institutional sources, including the US Defense Modeling and Simulation Office (DMSO) [DMSO, 00], the NATO [NATO, 98], the International Test Operation Procedure (ITOP) on V&V [ITOP, 04], and the AFDRG MEVAS project [Jacquart et al., 03], as well as commonly known scientific contributions, such as [Shannon, 75] and [Zeigler et al., 00].

The follow-on project of REVVA, EUROPA-111-104 (“REVVA2”), was launched in January 2006, and will run for three years. The results from REVVA2 are expected to constitute a generic VV&A methodology, and will be submitted to the SISO (Simulation Interoperability Standards Organization) as a VV&A methodology standard proposal. Six nations are participating in this project: France (lead nation, ONERA), Canada (DND SECO), Denmark (UNI-C), the Netherlands (TNO), Sweden (FOI and FMV) and UK (SE Validation). Currently, the REVVA2 consortium is taking part actively in the Product Development Group “Generic Methodology V&V” within SISO.

This section gives an overview of the current results of the REVVA/REVVA2 projects. In the sequel, the REVVA/REVVA2 projects will be referred to simply as “REVVA”, and their results as the “REVVA methodology”. Those interested in more information concerning the REVVA methodology are referred to [PROSPEC, 04] and [METHGU2, 04].

2.1 Basic terminology

Below the most basic terms within Verification and Validation (V&V) are presented as they are defined by REVVA. The definitions introduce an important distinction among *properties* of products, i.e. correctness and validity, and the *process*, i.e. verification and validation, to perceive these properties:

- Correctness: The property of a simulation model to comply with formal rules and bodies of reference information for its content and representation, and for the transformation into another representation.
- Validity: The property of a simulation model to have, within a specific experimental frame, a behaviour which is indistinguishable from the behaviour of the System of Interest.
- Verification: The process which is used to construct, under a set of time, cost, skills, and organizational constraints a justified belief about model correctness.
- Validation: The process which is used to construct, under a set of time, cost, skills, and organizational constraints a justified belief about model validity.

This separation between properties and processes stresses the current situation that V&V cannot guarantee absolute correctness and validity.

[2] This section is directly based on [PROSPEC, 04] and [Brade et al., 05] by reusing the texts from there with only minor changes and updates.

2.2 The “three pillars”

The REVVA methodology is based on the assumption that the most fundamental aspects that should be addressed by a VV&A methodology are process, products and organizations (“the three pillars”).

- The *organizations*. REVVA assumes that the potential value of an M&S VV&A effort strongly depends on the organizational context. In general, the quality of the organization and the way of allocating and sharing the work and responsibilities is of primary importance, which involves different groups with different, sometimes conflicting, interests. See section 2.3.
- The *process*, which directs the flow of activities and products during VV&A. The *REVVA Generic Process* is a stand-alone VV&A process which can be mapped to standard modeling processes via the M&S intermediate and final products made available for VV&A. See section 2.4.
- The *products*, which document the findings of the VV&A effort. This pillar is mainly built out of Items of Evidence, which are the basic results of the application of V&V Techniques. It is structured according to the semi-formalized Acceptability Criteria (documented in the Target of Acceptance, ToA) and the chosen V&V approach (Target of Verification and Validation). The basic results are integrated for the acceptance decision into an overall picture. See section 2.4.

This three-pillars model is a meta-description of the REVVA methodology. It captures the dependencies of and flow of information between the methodology components. It is expected that making these relationships explicit should be beneficial for the comprehensiveness, focus and balance of the VV&A project.

2.3 Organisations: Parties and roles

The pillar “organizations” is implemented in terms of parties and roles in the REVVA methodology. Groups with different interests, including those who are going to acquire a simulation model or simulation results (and are likely to pay for it), and those who deliver the requested M&S product, are distinguished. These interest groups are called *parties*. A party is assumed to be an organization or organizational unit. With the situation that somebody provides a simulation model or simulation results, which will be used by somebody else, there exists a “customer-supplier relationship”:

- *Customer*: A customer is an organization or organizational unit which plans to use or is using an M&S product (such as a SEM, simulation results, or data) developed by another party.
- *Supplier*: The supplier is an organization or organizational unit which provides the M&S product.

A relationship of trust between the customer and the supplier is desirable, but it must be always kept in mind that the supplier is trying to sell something to the customer, with all its implications. Thus, the REVVA methodology introduces the:

- *3rd Party VV&A Agent*: The 3rd Party VV&A Agent is an organization or organizational unit external of the customer and the supplier parties. Its degree of independence is assessed based on managerial, technical, and financial factors.
- *Acceptance Authority*: The Acceptance Authority is an organization or organizational unit external of both the customer and the supplier parties, officially entitled to accept M&S products, and trusted by the customer. Its degree of independence is assessed based on managerial, technical, and financial factors.

A *role* is characterized by the skills required to accomplish a particular task or set of tasks, and the responsibilities that are taken. The roles are played by actors from the above parties. The decision, which party an actor comes from, must be made carefully and deliberately. *VV&A core roles* are directly involved in the VV&A endeavour by using, planning, conducting, evaluating, or assessing the substantial VV&A work:

- The *Contextual User* defines the contextual objective. It is assumed that the Contextual User always is in the customer party.
- The *Acceptance Leader* is a user representative (trusted by the Contextual User), who is responsible for the assessment of the M&S product. The person in this role also finally judges the success or failure of the V&V effort.
- The *V&V Leader* knows approaches to V&V, techniques, and tools. This role is responsible for developing an appropriate V&V approach to substantiate the Acceptability Criteria (AC) with the information about System of Interest and simulation model available.
- *The V&V Executioners* is a composite of roles including Simulation Model Operators, System Analysts & Subject Matter Experts, M&S Experts, and HW/SW Engineers. It consists of a number of actors playing several roles that actually implement the analysis and test activities required to provide the Items of Evidence specified by the V&V Leader.

Affected roles are also identified. These roles are not directly involved in the technical planning and implementation of VV&A. Often they are decision makers outside of the process, are responsible for the smooth organizational flow of the VV&A effort, and control the flow of information among all parties involved. They are *M&S Promoter*, *M&S Sponsor*, *M&S Project Manager* and *VV&A Project Manager*.

Whether an actor or group of actors is appropriate to play a particular role depends on organizational aspects, including the desired degree of independence and required transfer of information, and on her/his educational back-ground and experience. The REVVA methodology distinguishes:

- Dependent V&V (DV&V): The V&V is conducted by the M&S supplier according to the customer’s V&V requirements (i.e., the actors for V&V Leader and V&V Executioners are members of the supplier party), and accepted “as is” by the customer.
- Independent Assessment (IA): The V&V work is conducted by the M&S supplier, but is assessed by an independent Acceptance Leader (from an independent 3rd Party) trusted by the customer.
- Independent V&V (IV&V): V&V activities are planned and conducted independently from both the supplier and the customer by the independent 3rd Party VV&A Agent.

2.4 The REVVA Generic Process (RGP) and the associated products

This section presents the RGP, the products produced as output from the process, and the M&S products required as input to the process. The RGP is shown in Figure 2-1. The process consists of phases (described by boxes) and products (ellipses). Each phase description contains a summary of activities, lists the input and output products, and points out the involved roles and their type of involvement. The REVVA Generic Process is no waterfall process, but iterative. For details the readers are referred to [PROSPEC, 04].

Develop ToA (phase 1): Based on the intended purpose of model use, a detailed set of AC is developed in such a manner that passing the AC implies fitness for purpose. All AC and the rationale for their derivation are recorded as the “Target of Acceptance” (ToA). AC should be prioritized. For simulation based endeavours with a low impact on real world decisions or actions, some superficial indicators that the AC are passed may be sufficient, while safety critical aspects might require an unmistakable proof.

Target of Acceptance (product): The Target of Acceptance (ToA) contains a precise specification of the AC and the rationale for their derivation from the intended purpose, and documents “what needs to be demonstrated” during the V&V effort. On top of a refinement hierarchy stands the vague intended purpose, which is refined into a set of sub-purposes, which again is decomposed, until AC related to the M&S product’s correctness and validity can be derived directly from the lowest sub-purposes. The set of AC does not imply any methods or techniques concerning how to assess them.

Acquire Information (phase 2): Under consideration of the intended purpose of model use and the detailed AC (documented in the ToA), knowledge about the System of Interest (SoI), its structure and behaviour, its subsystems and their structure and behaviour, or related systems is collected and filed (in related work this body of real world knowledge is referred to as “referent”).

Model information and system knowledge (product): This information will be used as foundation of the approach to demonstrate the model’s correctness and validity. The product identifies all sources of information, knowledge and all bodies of information and knowledge that are available or will become available during the V&V effort. The acquired information and knowledge about both the M&S product and the SoI is ideally stored in (an) appropriate remotely and securely accessible data base(s).

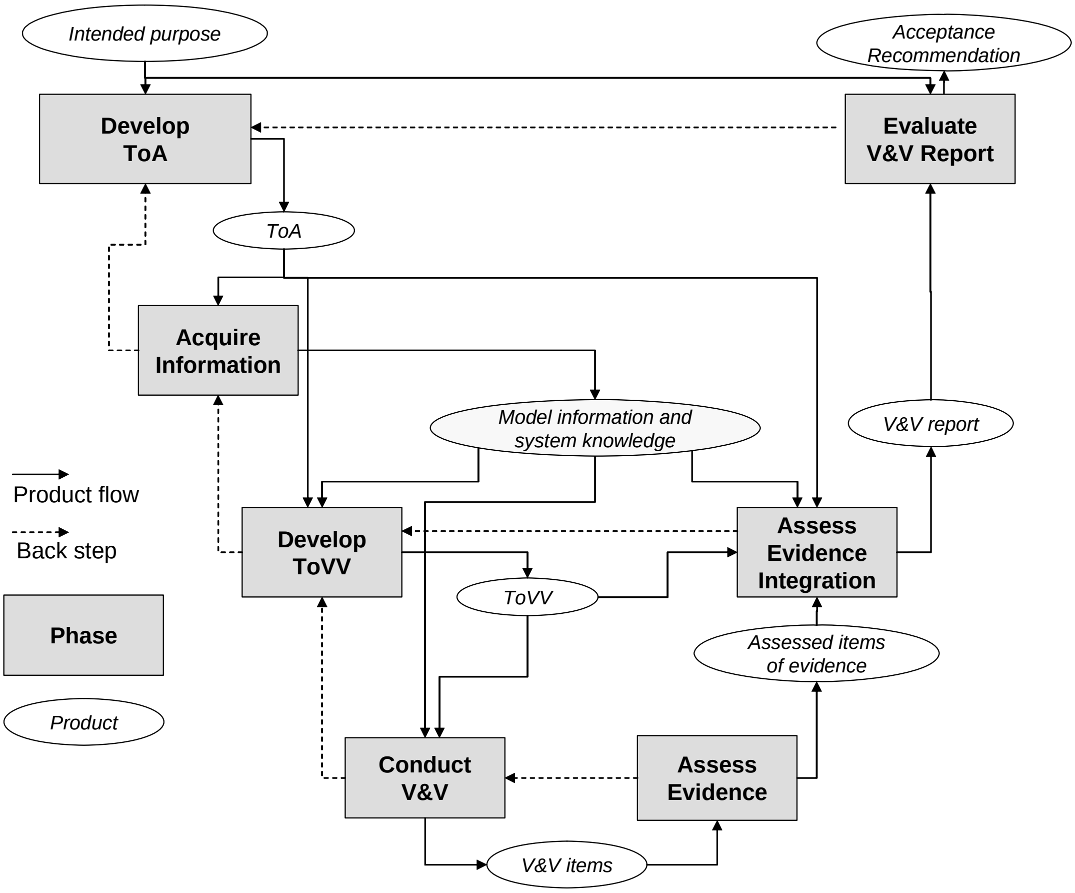


Figure 1. The REVVA Generic Process and Products

Develop ToVV (phase 3): For each AC a rationale is developed, which points out how with the information at hand and the available technical means it can be demonstrated that the AC is passed or failed. To substantiate that the AC is met becomes a V&V Objective. Developing the ToVV usually includes the decomposition of a V&V Objective into more easily assessable V&V sub-objectives.

Target of Verification and Validation (product): The Target of V&V (ToVV) documents the approach taken to the substantiation of the AC. It elaborates on "how to demonstrate that the AC are passed or failed", identifies the Items of Evidence (IoE) required to substantiate the AC contained in the ToA, and documents the rationale for the necessity and sufficiency of these IoE. The rationale for this decomposition includes the justification, why passing the lower V&V sub-objectives also implies passing the AC from which they were derived. Besides the required information within the IoE, the ToVV also identifies their individual desired probative forces needed to consider them "strong enough".

Conduct V&V (phase 4): V&V is conducted to provide the V&V items required by the ToVV. If, due to, e.g., missing or insufficient information about the model, missing knowledge about the SoI, or unavailability of the required tools, a particular IoE cannot be acquired, or if an elementary V&V objective is demonstrated to be failed, a step back to "Develop ToVV" is made.

V&V items (product): Each test result, analysis report, or proof outcome is documented as V&V Item, which as a set, constitute the "atomic building blocks" of V&V. A V&V Item consists of some piece of information about the simulation model, the evaluation objective, reference information, an evaluation technique, and the evaluation result. For validation, the reference information consists of knowledge about the SoI. For verification, the reference information consists of, e.g., representation rules, model information in a different representation form, or formalism. V&V Items have different probative forces, depending on the method or technique used for their creation, and the reference information or knowledge used.

Assess Evidence (phase 5): The key issue of this phase is to assess the probative force of the V&V items, to accept the individual V&V items as IoE, or to reject them. If the probative force of an IoE is considered to be unacceptably low, the IoE needs to be strengthened by repeated conduction of V&V activities or discarded. Otherwise, the IoE is added to the evidence pool, which its perceived probative force annotated. The probative force of each individual IoE is assessed based on the repeatability of the associated V&V activity. The probative force of an IoE is considered to be high, if the V&V result is reproducible, independently from its subjective elements (human beings). It is considered to be low, if it strongly depends on its subjective elements and its various results depend on the different individuals involved.

Items of Evidence (product): The Items of Evidence (IoE) document the individual executions of single V&V techniques and their outcomes, as conducted or acquired by the V&V Executioners. The assessed IoE includes (in addition to the information contained in the V&V item from which the Item of Evidence originates) the assessment statement, and a judgment of its probative force.

Assess Evidence Integration (phase 6): A single IoE will usually not allow the conclusion that a particular AC is passed, but several IoE are assembled according to the (most recent version of the) ToVV. The key issue of this phase is to build and accept or reject the rationale of supporting the AC with the available IoE. Under reconsideration of the ToA, the assembly of the evidence is reviewed and it is judged how sufficiently the evidence substantiates that the AC are passed (convincing force). An AC is considered to be completely covered, if the rationale for the derivation of directly succeeding V&V sub-objectives makes clear that meeting the V&V sub-objectives automatically implies meeting the parent AC, too. If the available evidence leaves unacceptable gaps or loopholes for the substantiation of the AC, the ToVV needs to be adjusted and the additional V&V activities conducted to provide the missing IoE.

V&V report (product): The IoE assembled and integrated by the V&V Leader to substantiate the AC in the ToA according to the most recent version of the ToVV, build the substance of the V&V

report. The V&V report links the rationale why the referenced IoE substantiate the claim that the AC are passed with the IoE made available.

Evaluate V&V Report (phase 7): Based on the probative force of the evidence, the convincing force of the ToVV, and the selection of AC as motivated in the ToA (all documented in the V&V report), the residual uncertainty associated with the statement that the M&S product actually is fit for its intended purpose is estimated. If the residual uncertainty is considered to be too high, either the intended use must be modified in such a manner that invalid simulation results have a less critical impact, or the V&V effort must be partially repeated with an extended ToA. The level of residual uncertainty needs to be identified for each AC and each relevant set of AC individually. While for particular AC a high degree of uncertainty may be acceptable (criteria which may be failed without serious consequences), for others only very low uncertainty may be acceptable (criteria whose failure will have serious impact).

If no disproving evidence has been acquired, if the affirmative evidence is considered to be “strong enough”, and if the strategy according to which the affirmative evidence is assembled to substantiate the claim that the AC are met is considered to be “sufficiently convincing”, then the M&S product is perceived as correct with respect to all relevant specifications and constraints, and as valid for its intended purpose (as represented by the ToA) with sufficiently low residual uncertainty. To prepare a responsible acceptance or rejection decision, an upper bound for this residual uncertainty is estimated.

Acceptance Recommendation (product): The final recommendation whether to accept or reject the M&S product for its intended use, considering the uncertainty that is left even after V&V was successfully conducted, is documented in form of the acceptance recommendation. The acceptance recommendation confirms that the acceptability for the intended purpose is demonstrated by the IoE gathered to substantiate the AC, and states a reasonable degree of confidence in this confirmation.

3 CGF and HBR

Bakground

The term Computer Generated Forces, CGF, was coined in the 1980'ies and describes computational models of behaviour that represents both individual actors and groups of actors. The goal of CGF development is to populate simulated worlds with computer generated actors that in different situations and circumstances to a varying degree behave and act as people do in the real world. CGF development can have the goal of representing how humans really act, but CGFs can also provide good training value with less elaborate behaviour. Since eight to ten years the CGF term is by and by generalized to HBR, Human Behaviour Representation, since many of the problems and solutions of development and deployment are similar and cover many domains and applications besides the training of military forces. The terms computer generated forces (CGF), semi-automated forces (SAF and SAFOR), synthetic forces, automated forces (AFOR) and command forces (CFOR) all refer to different forms of HBRs.

In applications ranging from educational instruction to personnel training, human behaviour representations (HBRs) can be designed to efficiently carry out an array of vital tasks (e.g., performance assessment, design of crew stations and interfaces, diagnosis of training needs, intelligent tutoring).

To develop CGF with realistic human behaviour is a very big challenge. [Harmon et al.]

HBRs are unique among other complex simulations. At first blush, they appear distinguished from the other parts of a simulation by their

- Very high inherent complexity,
- Numerous nonlinear relationships all interacting chaotically over many different orders of magnitude, and
- Complex coupling with other parts of a simulation system.

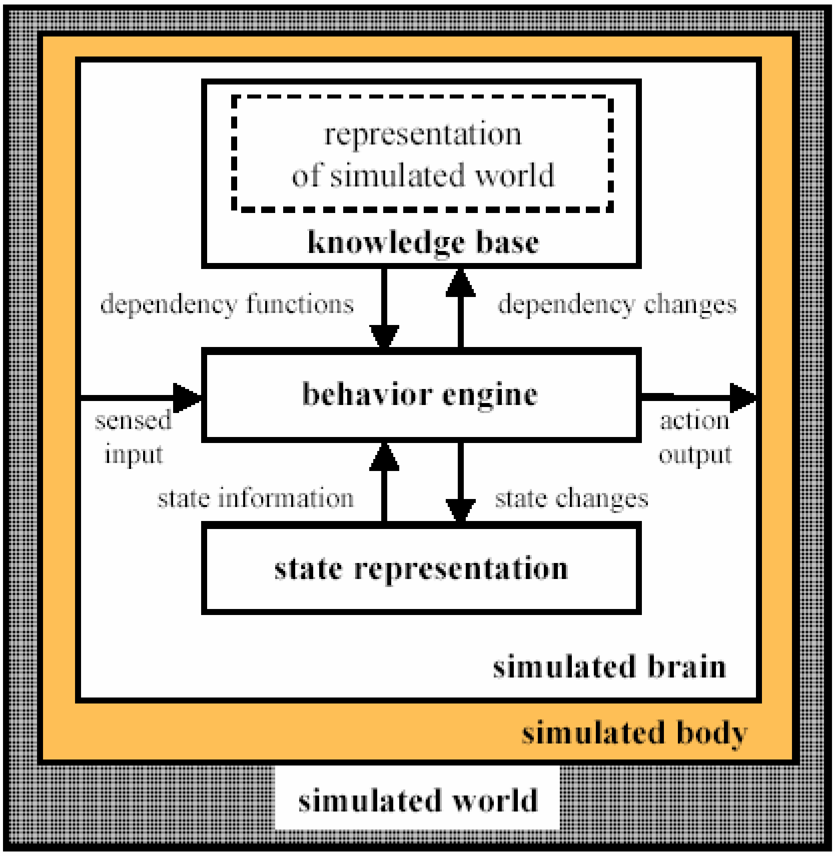
All simulations must consider abstractions. The profound question when using HBRs must be to what extent the model should correspond to the real world. This is a matter of validity, what are the abilities, the capabilities and the constraints. What are the requirements of correctness and detail?

HBRs for different applications show a wide span of functional requirements and this span range from simple behavioural functions to more process demanding implemented functions. This separation or rather continuum of abilities is essential to specify because this put highly different demands on the validation efforts to come.

Description of HBR Nature ([Harmon et al.])

As mentioned above, all HBRs model the behaviour of people at some level of abstraction. So, HBRs can model any combination of the many different facets of human behaviour including

- Ability to reason (e.g., knowledge based systems)
- Ability to change the environment (e.g., operating equipment)
- Responds to comfort and discomfort (e.g., environmental safety)
- Susceptibility to injury and illness (e.g., injury models)
- Emotional responses (e.g., affective models)
- Ability to communicate with other humans
- Abilities to sense the environment (e.g., vision models)
- Physical capabilities and limitations



Figur 2. Description of a Generic Human Behaviour Representation [Harmon et al]

Figure 2, the HBR canonical model [Harmon et al], depicts the basic components of a simulation of the neurologically-related human behaviour, considered by many as generating the most interesting parts of human behaviour. In this model, the knowledge base consists of the executable dependencies needed to create the internal state representation from sensory input and to respond to that state. The knowledge base also includes the decision functions that determine when and which of those dependencies should be executed to achieve goals at any particular time or combination of stimuli. The behaviour engine chooses the dependencies from the knowledge base appropriate to the current state and executes those dependencies to modify the internal state representation or to generate the actions to achieve the HBR's goals. The state representation depicts the HBR's dynamic assessment of both the internal and external world state including all goals.

Architectures

Architecturs or frameworks for developing cognitive process models have been built since the 1980'ies. They are the tools for being able to produce, within a reasonable time, models that are built on adopted theories on cognition. There are several reviews of such architectures that will not be included in this report. [Gluck & Pew 05] has a list of twenty such architectures referenced in Web sites as well as references to other authors that has been exercising and surveyed all of them. Such architectures sometimes are called pshycological or cognitive or human representation or integrative or hybrid.

Requirements specification.

DMSO, in its VV&A Recommended Practice Guide [RPG, Build 2.5] under Special topics, points out the most important area on specification requirements and their relation to the task of validation. The original intent or goal definition for a model or application is the base for the requirements. The M&S expert must decide which abilities should be modelled as a consequence from that. One requirement can be that these abilities and capabilities should be measurable. In the FEDEP process requirement specifications are decided in the second phase, Conceptual Analysis, These specifications are on one hand the base for the model design and on the other hand the base for validation efforts as validation and acceptance criteria.

The Requirements are sometimes separated into three categories:

- Functional requirements
- Requirements on fidelity
- Implementation requirements

The requirements on fidelity are the most difficult to analyse, choose and effectuate.
This has a direct coupling to the different types of validity that is discussed in later chapters.
A general, though rough, proposal of classification is found in [NATO SAS017, 2001] (Human Behaviour Representation, RTO TECHNICAL REPORT 47), with fidelity broken down as follows:

- Very Low – Spreadsheet analyses.
- Low – Black box model that captures inputs and outputs, but makes no assumptions about internal cognitive processes. Generally mathematical models with Input/Output face validity – intellective models. Sometimes called "**performance**" models.
- Medium – High fidelity black box model, involving some explicit process modelling
- High – Emulation or **Process** model describing the details of the transition from inputs to outputs. Sometimes called "**functional**" models.

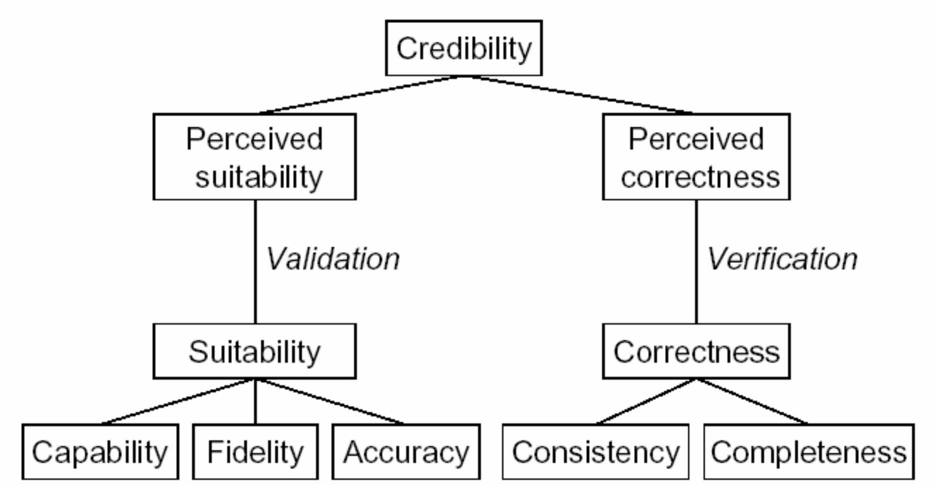
4 VV&A

In today's M&S community, the terms "verification" and "validation" are usually mentioned as part of the triplet "VV&A", with the "A" standing for "accreditation". Accreditation is a bureaucratic act, during which a model or model results officially are declared as acceptable for a specific intended purpose [Defense Modeling and Simulation Office 1996 and 2000]. Accreditation should be exclusively based on the credibility of the model.

4.1 Verification and Validation of Models and Simulation Results

The purpose of M&S is to represent a real system in order to draw conclusions about the real system by experimentation with a model. Thus, the direct correlation between the model, an intended purpose of model use, and a clearly identified real system are among the key characteristics of simulation. The term "simulation" implies a claim to represent the behavior of a real system as it is or as it could be. The direct association to a real system distinguishes computer-based simulation from, e.g., computer games. As decisions that heavily impact the real world rely increasingly on models or simulation results, the more important their *correctness* and *suitability* becomes. Suitability refers to the concepts of *capability*, *fidelity*, and *accuracy*, while correctness refers to *consistency* and *completeness.* The growing role of M&S implies that measures must be taken to ensure the correctness and suitability of models and simulation results. As neither suitability nor correctness can be proven in most cases, the *credibility* of a model or simulation results is of major importance. The credibility of a model is based on the *perceived suitability* and the *perceived correctness* of all intermediate products created during model development. The correctness and suitability of simulation results require correctness and suitability of the model and its embedded data, but also suitable and correct runtime input data and use or operation of the model. *Verification* and *validation* aim to increase the credibility of models and simulation results by providing *evidence* and *indication* of correctness and suitability. The dependencies between the terms introduced above are illustrated in Figure 3, [Brade 2004].

Figure 3: Dependencies between V&V related terms



4.1.1 Definitions

Beside the definitions given in chapter 2.1 you can easily find following alternative definitions:

Model verification is the process of demonstrating that a model is correctly represented and was transformed correctly from one representation form into another, according to all transformation and representation rules, requirements, and constraints.

Model validation is the process of demonstrating that a model and its behaviour are suitable representations of the real system and its behaviour with respect to an intended purpose of model application.

Definitions of verification and validation

The following definitions are drawn from DMSO [DoD 5000.59-M]

Verification:

1. The process of determining that a model implementation and its associated data accurately represent the developer's conceptual description and specifications.
2. The process of determining that a model or simulation faithfully represents the developer's conceptual description and specifications. Verification evaluates the extent to which the model or simulation has been developed using sound and established software and system engineering techniques.

Internal consistency:

The process of determining that a model implementation and its associated data accurately represent the developer’s conceptual description and specifications.

Validation:

The process of determining the degree to which a model or simulation is a faithful representation of the real world from the perspective of the intended uses of that model or simulation

*Fidelity***:**

The accuracy of the representation when compared to the real world.

4.2 Distinction between validating a model and an architecture

It is important to demonstrate that a cognitive architecture will implement processes predictably across all models developed within that architecture. However, it is not possible to guarantee that every model built in an architecture will be an accurate representation of its referent in the real world.

The bottom line is that different architectures impose different levels of constraints, and there is no architecture that perfectly constrains all model development to only valid models.

5 VV&A of HBR

5.1 The challenge

The validation of HBRs has posed particularly vexing problems since their very first applications. HBRs can easily create extremely large, convoluted and non-linear behaviour spaces. Even simple HBRs can differentiate thousands of situation conditions and produce hundreds of responses to those conditions. Highly non-linear relationships between perceived situations and derived responses to those situations are commonplace. This nonlinearity means that the behaviour observed, and perhaps validated, for one set of conditions cannot be generalized for another set even though the differences between conditions may be small. HBR knowledge bases represent executable, as well as state, information and may contain the same volume of information as moderately complex environmental representations. Small situation changes over time often create wildly different responses in the same system. Validation of HBRs, even for simple tasks, can prove extremely difficult because of the large number of behavioural paths that must be explored for any given application. Nonlinearities and complex couplings prevent the reliable use of sampling and extrapolation techniques in results testing. The lack of well established techniques and tools to support HBR validation further exacerbates the difficulty of these problems. Therefore, a major challenge for research supporting the validation of models of human behaviour lies in the development of a strong methodology and accompanying toolset for the HBR-community to use in validation. (Cited from [NATO RTO SAS17 01].)

5.2 Referents for HBR Validation

Referent: A codified body of knowledge about a thing being simulated.

5.2.1 What purpose does a referent for an HBR serve?

A simulation’s referent defines the standard against which to measure its representational capabilities to determine the accuracy of its representations.

5.2.2 What referents exist for validating HBRs?

A referent represents the total collection of knowledge about a particular subject, in this case, human behaviour under various circumstances. Referents for HBRs can come from:

- SMEs (Subject Matter Experts
- empirical observations or experimental data from actual operations
- validated models of various aspects of human behaviour
- validated models of the physiological processes underlying human behaviour
- validated models of sociological phenomena (useful particularly for modelling groups of people)
- . validated simulations of human behaviour

5.2.3 Levels of referents for HBR validation

Like models of complex physical processes, HBR models can be validated at many different levels of abstraction. The list below illustrates six levels of model correspondence for HBRs.

Levels of HBR Correspondence and Referents

- Domain Correspondence
- Sociological Correspondence
- Psychological Correspondence
- Physiological Correspondence

- Computational Correspondence
- Physical Correspondence

A HBR that has correspondence at all six levels best approximates human behaviour for all purposes. Most purposes may only require correspondence in one or two of these areas.
(http://vva.dmso.mil/Special_Topics/HBR-Validation/hbr-validation-pr.pdf)

These issues are more discussed later on in 5.4.3.

5.2.3.1 Perspectives on validation that can be found in the psychological literature.

Psychologists have been testing theories and hypotheses about human behaviour for a long time. Cronbach & Meehl [Cronbach & Meehl, 1955] focuses on validation of psychological tests, how we can know that a person's score on a test has any meaning. Within this context, they present a taxonomy of types of validity:
Criterion validity – the extent to which a person's score on a test predicts that person's performance or score on some other independent measure of interest
Content validity – the extent to which the test items form a representative sampling of the potential universe of all relevant content
Construct validity – a measure of the extent to which a test score is an accurate reflection of some underlying psychological trait or characteristic of the test taker.

Cook & Campbell (1979) discuss the validation of an experiment or experimental design, how we can know that an experimental result has any meaning and we can have confidence in the accuracy of the inferences about causal relationships between variables that are based on the results of an experiment.

Types of validity:
Statistical conclusion validity – the extent to which statistical requirements (sample size e.g.) are met in the experimental design
Internal validity – the extent to which causal relationships have been accurately defined among manipulated variables
Construct validity – the extent to which those same causal relationships generalize to the underlying psychological traits of interest
External validity – the extent to which the identified causal relationships generalize across populations and environmental conditions.

Both definitions of construct validity are based on the premise that humans have psychological traits that cannot be measured or studied directly, but can only come to be understood through inferences based on imperfect indicators. A construct HBR would be the one in which the knowledge base and behaviour engine implemented in the model correspond to the knowledge structures and cognitive and psychomotor processes of the person or people being modelled.

A second important similarity is that both taxonomies treat construct validity as a continuum rather than an all-or-nothing judgement. There is no way to prove that something is valid. Instead the relevant community must establish a threshold of acceptability or sufficiency that individual efforts can be compared against. (This section is cited from [Gluck & Pew 05].)

Another reference that discusses types of validity is Major James Denford in his Experimentation Guide, [James Denford 2005], point 37:
Concepts of Reliability and Validity. Reliability is the consistency with which the measuring instrument performs. This is seen as the ability to repeat or replicate the same study obtaining similar results. Validity, on the other hand, is concerned with whether a variable is the underlying cause of variation in a phenomenon. Leedy identified six types of validity: face, criterion, content,

construct, internal and external. Different research methods and approaches to those methods have different strengths and weaknesses in reliability and validity, with the researcher required to select the appropriate method based on research goals and an understanding of the limitations of various methods.

a. Face validity depends on the researcher's best judgment of whether the instrument is measuring what it is intended to and if the sample is representative of the trait being studied;
b. Criterion validity is based on an instrument or scale having an empirical association with some criterion or standard;
c. Content validity is the accuracy with which an instrument measures the factors being studied;
d. Constructs are hypothesized entities that cannot be measured directly; hence construct validity refers to the degree to which the construct itself is indirectly measured by the scale or instrument;
e. Internal validity is a freedom from bias in establishing the causal relationship between variables of interest; and
f. External reliability deals with the generalizability of the conclusions beyond the bounds of the selected sample.

This is definitions (ontology) of types of validity from a primarily psychological point of view. Some of them e. g. application validity is discussed more later on in this paper. But before that we will se what is common practice today what technology resources can be used.

5.3 Current practice and deficiencies

Current HBR validation practice relies nearly exclusively upon *face validation* by subject matter experts (SMEs). During this process an SME drives the HBR through the scenario space by issuing commands or changing the stimulating situation, observing the resultant behaviour, and determines, often qualitatively, whether that behaviour meets an application’s requirements for realism. In some cases, the SME only watches the HBR within the context of a much more complex simulation execution to assess its validity without direct interaction. Behaviour anomalies may require modifying the scenario (the easiest of options), the knowledge base (the next easiest option) or the behaviour generation mechanisms (the hardest option) then repeating the process. Some feel that relying primarily or exclusively upon face validation minimizes or entirely avoids the much more costly and complex validation of models or knowledge bases. Regrettably, face validation is also the least reliable, least complete and, therefore, weakest form of HBR validation. Reliable and sufficiently complete SME characterization of HBR behaviour spaces is only possible with the simplest of representations. This current practice will become totally inadequate as the HBRs include more and more realistic (i.e., higher fidelity) behaviour. Fortunately, some technology exists to overcome these limitations to some degree. [NATO RTO SAS17 01].

5.4 What is currently available: technology resources

This chapter focuses on this issue and attempt to incorporate insights from several disciplines, including software engineering, mathematics, statistics and psychology.

5.4.1 Tools and techniques

Tools for KBS, knowledge based systems, assess along the dimensions of completeness, consistency or coherence, and redundancy.

There are examples of tools even for validating software and, in particular, simulations. But are these good enough? HBR validation is the most challenging task because HBRs actually often are made up of two sets of computer programs: a behaviour engine and a knowledge base.

Thus we need to look beyond the tools and techniques available within the software engineering community to validate HBRs.

Establishing the validity of human behaviour representations (HBRs) involves comparing a model or simulation against the available referents to determine if its behaviour faithfully represents that of actual humans sufficiently enough for a particular application. Without proper validation, no HBR can be confidently applied to any problem with predictable results. Their inherent complexity makes HBRs the most difficult of any model to validate and no reliably repeatable techniques to perform that task. However, technology does exist that can improve that situation. Application of this technology can significantly improve results testing and make HBR validation cost effective for a wide variety of applications.

Two primary technology resources presently exist to support HBR validation: knowledge based system (KBS) verification, validation, evaluation and testing (VVE&T) technology, and human behaviour science [NATO RTO SAS17 01].

5.4.2 KBS VVE&T Technology

Significant investment has gone into the development of KBSs to perform a variety of expert functions including diagnosis, decision support and automatic control. KBSs have seen application in such critical areas as flight control, financial management, and disease diagnosis and treatment recommendation. These applications, together with the enormous success of augmenting human expertise with machine intelligence, have driven the development of technology to measure and improve the validity of KBS behaviour. To some extent, HBRs are a class of KBS so the technology supporting the VVE&T of KBSs presents a tremendous and, heretofore, untapped resource for HBR validation.

The literature describing KBS VVE&T theory seems comparatively limited, less than 3% of the literature surveyed, but it covers all of the important problems including data selection, verification, validation and testing. Further, the challenging task of developing the theory that underlies the behaviour of KBSs has only recently seen some promising advances. Until a comprehensive and consistent theory of intelligent systems exists, the theory supporting VVE&T of KBSs will likely remain as loosely coupled conceptual islands. On the other hand, a myriad of verification and validation (V&V) techniques and tools has been proposed, developed and tested. The literature surveyed discusses 41 different VVE&T techniques that address logic, optimisation, classification, transformations, graph theory, empirical techniques, heuristic techniques, formal methods, optimisation, modelling and simulation. These techniques addressed knowledge base integration (examining completeness/coverage, consistency/coherence and redundancy), various knowledge conditions (including incomplete, multi-level, modular, uncertain and incorrect knowledge), numerous specific representations (e.g., nonmonotonic logic, case-based reasoning, tabular representations, equations, weighted rules, control/meta-knowledge and dynamic properties) and architectures (e.g., blackboard systems, expert system shells, and multi-agent systems), and various aspects of verification and validation processes (covering automatic refinement, knowledge base verification, subjective criteria, large knowledge bases and wide domains). The survey uncovered 60 tools that support all different aspects of KBS VVE&T. These varied from single tools with limited capabilities and associated with specific expert systems to rich integrated tool sets that apply to any KBS written in a particular programming language (e.g., PROLOG or OPS-5) or using a specific expert system shell. This literature also revealed an enormous amount of experience in VVE&T of KBSs with 115 different references for diverse applications. By far, most of this experience related to medical applications where the results from any KBS can have life threatening consequences. [NATO RTO SAS17 01].

5.4.3 Human Behaviour Science

Validating the models of a simulation involves comparing the characteristics of those model abstractions with the most detailed knowledge of the modelled behaviour available, i.e., the referent. This comparison identifies where the models coincide with and deviate from the referents. This information determines how well the models meet the fitness requirements of specific applications. HBR referents could come from experimental observations and theoretical approximations of human behaviour as well as SMEs. Considerable literature exists in these areas thus providing a rich source of referent knowledge.

Domain referents can come from experimental data. Some quantitative experimental data exists on actual human performance in various battlefield situations (e.g., data from instrumented ranges) and from humans performing very specific cognitive tasks under controlled conditions. Regrettably, much domain specific experimental data is very sparse and applies only to narrow situations. Often the experimental conditions for data collection are very poorly controlled and characterized. These source data problems weaken any validation done against them. However, as the technology develops, better experimental data for different domains will become more widely accessible.

As with psychological validation, a rich body of sociological knowledge exists from which referents and tests for team, group and organizational behaviour can be drawn. This knowledge includes both models describing sociological phenomena and experimental observations. Sociological experiments also provide well established experimental protocols to support the design of tests that characterize the correspondence of team, group and organization simulations with their referents. Current sociological knowledge permits the testing of behaviour manifested by groups as well as of the interaction dynamics between the members of those groups. Testing *sociological validity* is particularly important with simulations of human groups cooperating to perform some task and may not be necessary when representing the actions of isolated individuals.

A vast body of knowledge exists about human psychology. This knowledge includes numerous abstract models of many different aspects of human behaviour as well as an enormous volume of published experimental data on actual human performance under different circumstances of interest that validate these theories to some degree. Experimental data can completely establish a referent or augment that created by psychological models. Testing *psychological correspondence* creates stronger validation than domain correspondence testing alone because of its linkage to the underlying psychological phenomena. This linkage to the founding phenomenology means that the entire behaviour space of a simulation need not be explored because the psychological models inherently represent the nonlinearities associated with the behaviour they generate. Further, the experimental data from which referents are drawn has more likely been obtained under carefully controlled experimental conditions and is therefore more repeatable than that obtained from domain-specific experiments. Psychological correspondence testing enables validation of all of the HBR model components both as separate functions and as an integrated whole.

A considerable collection of experimental data and, recently developed, verifiable theory of neurophysiological processes begin to establish additional referents against which to compare HBR performance. HBRs have one significant advantage over the actual physiological systems from which this data originates; their detailed workings are easier to directly observe. Simulations that have *physiological correspondence* more likely behave like real people especially under conditions where non-neurological physiology contributes (e.g., fatigue and injury). This sort of evaluation is much closer to what has traditionally been done to validate physical system representations (i.e., non-human systems). In the past, this kind of validation has been difficult because the physiology of the human nervous system was not understood well enough to correlate physiological observations with cognitive behaviour except at extremely low levels (e.g., primitive vision). However, recent advances in non-invasive measurement techniques (e.g., MRI, PET) have improved our understanding of the linkage between cognitive behaviour and physiological observations and created a large repository of potential validation data. As this area of experiment improves, comparing these experimental results with HBR designs and performance will become easier and more meaningful. This form of correspondence testing might be particularly well suited for validating the effects of behaviour moderators, such as stress and emotion, and integrated models of human behaviour that incorporate such effects behaviour moderators. [NATO RTO SAS17 01].

5.4.4 Useful Techniques for collecting evidence

Techniques for collecting qualitative and quantitative evidence associated with assessing the validity of an HBR, inspired by the psychological notion of construct validity, is presented by Campbell and Bolton [Gluck & Pew 05]:

5.4.4.1 Qualitative or subjective evidence for model validity

The common way to achieve qualitative evidence is to ask subject matter experts (SMEs) to make judgements about the content and/or behaviour of the HBR. The extent to which a thing under consideration (measure, model etc) agrees with someone's common experience and intuition about how it should look and/or work is often referred to as *face validation.* There are several constraints and lessons learned how to form the process of collect assessments and opinions about the behaviour and look of the HBR from SMEs, known in literature of making tests and experiments and simulations with the help from SMEs. Among other are standardized, objective systematic, repeatable and independent processes and procedures involving questionnaires, checklists, structured interviews in advance.

Unfortunately, regardless of how carefully one collects these types of data, it is a well-documented fact that human judgements are prone to a number of limitations that make these data suspect. There is even a meaning among many psychologists that *face validation* is a false assurance of measurement for validation. The positive outcome may be that collecting validation evidences from SMEs give many good ideas for improvements….

5.4.4.2 Quantitative or "objective" evidence for model validity

The traditional way to collect quantitative evidence in the form of a statistical assessment of the similarity between an HBR's behaviour and a human's behaviour is done by collecting "objective" measurements.

Traditional statistical approach: Hypothesis testing. It can be shown that the use of traditional hypothesis testing procedures to compare model predictions to empirical data is inappropriate. This is due to that you cannot prove that the null hypothesis is true. That is what you have to do if you will prove that a model's predictions will be indistinguishable from empirically collected data.

An alternative statistical approach is Goodness-of-Fit Measures. Schunn and Wallace (2001) proposed that the goodness-of-fit between a model's predictions and empirical data should be assessed along two dimensions: trend consistency and exact match. However a high goodness-of-fit score from a comparison between your model and an appropriate set of data does not constitute strong evidence for the validity of your model. This is due to an effect in measurement theory known as *overfitting*. *Cross-validation* is a technique routinely used in the mathematical modelling community to assess the extent to which a model is overfitting.

Campbell and Bolton [Gluck & Pew 05] continues to discuss several limitations and weaknesses of Simple Goodness-of-Fit when applied to HBRs. However they also present that there are at least three ways to strengthen the validity evidence that can be accumulated by assessing the fit between model output and empirical data. There are the techniques of a) Model Comparisons, b) Pattern matching and c) A priori predictions.

5.5 Different types of validity.

5.5.1 Construct validity and cognitive models

Validaty is never definitely proven. Rather evidence is accumulated incrementally until it reaches a point that is considered satisfactory by some community.

From the work within the psychological community with the goal of establishing that a particular theory is construct valid which we can use as a base for model validation. In this context, construct validity means that the theory embodies an accurate description of the actual underlying processes that explain human behaviour, and that alternative theories can be disregarded.

But the question is: Do we always need that accurate desription of the actual underlying processes? Campbell and Bolton [Gluck & Pew 05] argues that is not always the case. They put the important questions: What is an appropriate goal for the developers and users of HBRs in applied, military settings? What does it mean to say that our models must be validated? Next section explains important means of "application validity".

5.5.2 Application validity: Assessing a model for its intended use

Concentrate on " … from the perspective of the intended uses". Different subcommunities, within the military community, have different goals with the use of HBRs. Military training community uses HBRs as synthetic adversaries in training simulators to increase the effectiveness of training activities, and thus increase trainees´ performance. The military acquisition community uses HBRs to evaluate candidate system designs and identify those designs that are likely to lead to the best human-system integration, and thus improve (human) performance. The military operational community uses HBRs as core components of decision-support systems (DSS), which are intended to provide explicit support to improve decision making in the field and thus performance. The point is that each community has an intended use, and an HBR can be assessed directly as to its ability to support that intended use.

Interestingly, there is at least some evidence that improving human performance does not necessarily require a construct valid model. This suggests that a model's capability to serve an applied goal (DMSO's definition of validity) is not necessarily equivalent to its construct validity. To distinguish these two types of validity, we use the term *application validity* to capture DMSO's meaning. Taking the "intended use" perspective serves two purposes: a) it bounds the scope of the validation problem; b) it provides insight into the activities, metrics, and measurement paradigm that could be used to demonstrate application validity.

5.5.2.1 Application: Training

The goal – to improve human performance – will already have been cast in:

1. terms of number of well-defined learning objectives,
2. learning activities will have been planned,
3. the conditions under which performance will be assessed will have been established, performance measurement techniques will be in place, and
4. criteria will have been set.

Demonstrating application validity of an HBR within a training community would require demonstrating that the incorporation of an HBR into a training system leads to some benefits, with improved human performance being the most obvious. There are many possible ways to measure this. Some common approaches include:

1. the average performance of a group of students increases,
2. the number of students who fail to met some minimum criteria decreases, and
3. the amount of time it takes the average student to reach a criterion decreases.

5.5.2.2 Application: System Design and Acquisition

An obvious use for HBR in this community is to support to simulation-based acquisition process. Here the application goal is to support the acquisition team (design engineers, lead system engineer, program manager, etc) as they develop, evaluate, modify, and finalize the design of a new military system.

Ultimately, assessing the application validity of an HBR for this community is not as straightforward as it is in the training community.

5.5.2.3 Application: Decision support System

A study assessing the application validity for this effort would appear similar to the validation effort for the training system described earlier.

Summary

The point is that the intended use of an HBR can provide insight into the types of evidence and assessment processes that should be used to validate that HBR.

5.6 V&V activities

Much of the content in this report is taken or cited from [Gluck & Pew 05]. In this book different authors describe and refer to the AMBR (Agent-Based Modeling and Behaviour Representation) project. As a matter of fact AMBR model comparison and their implications for the science of formal human behaviour representation are the focus of the book.

Methods and techniques, described in 5.4.4, for qualitative and quantitative assessments of models have been used in the AMBR project. Perhaps the strongest aspect among validation efforts of the AMBR project is that they even “pushed” the models to make predictions of performance under novel conditions prior to presenting the actual data to the modelers.

The authors of the HBR Validation chapter provide examples of many different types of evidence that could be collected to assess HBRs, as well as information about the capabilities of each to achieve two goals: that of providing supporting evidence for a claim of validity, and that of providing insight into ways in which a model could be improved.

Campbell and Bolton finally argue that the military community, in accordance with DMSO’s guidance, must also consider the specific goals for the use of the model(s) and allow the application to help shape the validation plan.

Gluck & Pew emphasize that it is not always the essential action for HBRs to behave exactly as humans do. They believe the criteria for success of models should be usefulness, not veridical representation of humans. Usefulness depends on many aspects of a simulation. If it is a training simulation, the usefulness is captured in measures of training effectiveness. If it is an evaluation associated with system acquisition, usefulness rests in the ability to discriminate real differences between alternative designs and so on.

5.7 Emerging Approches

Presently, HBR validation relies primarily upon behaviour testing by SMEs, the weakest and least reliable option. However, results testing will likely remain pivotal in HBR validation for the foreseeable future. This reality leads to the conclusion that improvements should be sought that can improve the coverage, reliability and strength of results testing. The application of existing and new validation technology should help target results testing efforts to focus upon those areas where problems are most likely to arise for an application and where automated testing cannot be trusted. [NATO RTO SAS17 01]

Existing KBS VVE&T technology represents a tremendous wealth of applicable theory, techniques, tools and experiences. Many of the techniques and tools discovered could be applied to future HBRs with appropriately selected implementation strategies. The largest amount of work related explicitly to expert and decision support systems. Most of this work specifically addressed KBSs using production rule knowledge representations. This focus upon rule representations could impact the validation of some existing HBRs that also use production rule representations.

Further, all of the VVE&T theory, techniques and tools apply only to the cognitive functions of HBRs and cannot be used for validation of the effects of behaviour moderators or performance limitations. Nonetheless, using KBS VVE&T techniques and tools could help to reduce the number of errors that exist in complex knowledge bases and integrated systems by automatically testing for completeness/coverage, consistency/coherence, and redundancy. Human testing of these knowledge bases is often immensely tedious and subject to error. Automated tools could also help to locate those areas in the behaviour spaces where behaviour is most complex and needs the most concentrated results testing. This would improve the value of test results and the confidence in the entire representation. KBS VVE&T technology can also contribute techniques and tools for automated test case generation that could support automated HBR testing. Such testing could further delineate those areas where SME expertise could be best applied.

All of the techniques drawn from behavioural science and applied to validating existing HBRs have significant limitations. As mentioned, testing *domain correspondence* requires unrealistic searches of very large and non-linear behaviour spaces. Testing *psychological and physiological correspondences* requires extensive validated models of psychological and physiological phenomena. While many comprehensive psychological models exist, relatively few of them have been applied to HBR validation, especially for simulation applications. Like the physiological models, many psychological models deal with very restricted behaviour spaces. These limitations prevent their useful application to HBRs representing behaviour for realistic situations. As psychological and physiological models become richer and more consistent, their utility for HBR validation will increase. A growing body of reliable and consistent referent data will soon make meaningful psychological and physiological correspondence testing possible and practical. As with models and simulations of physical systems, model correspondence testing must be done at several levels of abstraction. Only consistent results between these different levels can guarantee validity. *At this point, and probably forever, no single level of correspondence testing should be sufficient for any application.* Despite the current shortcomings, such correspondence testing can further circumvent the need to search large and non-linear behaviour spaces thereby reducing the cost and time required to reliably validate complex HBRs. When brought together, all of these resources (i.e., subject matter expert review, model correspondence testing, and KBS VVE&T tools and techniques) will dramatically improve the quality and reduce the cost of HBR validation through results testing.

These suggested advances can improve the coverage, accuracy and repeatability of results testing while reducing its cost and time to execute. *However, these goals can only be achieved through concerted cooperation between HBR developers, behavioural scientists and validation specialists.*

6 REVVA for CGF

The three pillars of REVVA are process, product and organisation.

The REVVA methodology is meant to be a generic method for VV&A of simulation models. From the above described about development and use of CGF and HBR and the plethora of definitions of validities it is an relevant question to ask whether the REVVA methodology is an appropriate method to use even for e g application validity purposes.

6.1 Process viewpoint

The discussion is sectioned for respectively the three most important types of validities.

For *construct validity* the interest is to have a simulation model, HBR model that is most applicable to the definition of validation given in REVVA. I refer to the term indistinguishable (behaviour from its referent in the real world). The behaviour from the HBR should be indistinguishable from the correspondence it is going to resemble. For this purpose it is natural to follow the REVVA process with the production of the appropriate documents given in the REVVA definition. In order to model cognitive functions, a great amount of knowledge from the field of cognitive science is necessary. Though several Cognitive architectures exist that will facilitate the development it will be a very complex and resource requiring procedure. As a matter of fact this should probably be more interesting for the psychologist community who is eager to produce, develop and test models of cognitive functions for primarily their own research and understanding of human cognition than it is for the M&S community that is more interested to develop the appropriate tools and models specific for military applications.

For the second type of validity, *content validity*, it is probably not appropriate to follow the REVVA process. The question here is to choose those functions and model contents that are defined to the extent that is decided to give behaviour god enough for an application with thoroughly given constraints in a clearly defined situation.

For the third type of validity, *application validity*, the model could be very simplified and therefore requires low validation effort. However, all the more effort must be made to recognise the requirements for the application and what these causes and even to be able to find some kind of measurements to support these.

On the other hand you can turn it around and ask what is useful or direct applicable in the REVVA process for those different kind of validities – and what else do you need?

6.2 Product viewpoint

The ToA document is the product of a generic break-down of the original intent of the application. This is a very useful method and knowledge that is suitable and recommended to use for every type of validity discussed above. When it comes to the ToVV it is obviously not so easy to produce in any of the three cases. When it comes to the rest of the documents there had to be some equivalents to the ones that is given in the REVVA methodology. For instance when it comes to application validity the only way to show that application validity exist would be to show that there is a positive result in e g training results if the CGF is used in a training application,

6.3 Organisation viewpoint

There is a common understanding within the HBR development community that to reach progress and make successful contributions of VV&A of HBR for the M&S area there is a need for common efforts from specialists in appropriate branches, sponsors, producers, consumers and controllers both for development of useful methodologies but also in actual development projects. The organisation proposed in the REVVA methodology seems very adequate for such a purpose.

6.4 Evaluation

[Harmon et al] are tabulating a few common myths about HBR validation. The final one expresses "Validating HBRs is too hard so why do it or even try to understand it." But in refutation to this they also say that "Good understanding of HBR validity can simplify the difficulty of abstracting the parts of human behaviour necessary to achieve a purpose" and "Like any validation task, a reasonably simple discipline can produce acceptable and cost effective results".

Concerning the VV&A methodological issues and difficulties the chairman in the REVVA-group expresses that we will have to separate the pure science oriented problems, the technical demand to model humans and their behaviours, from the methodological aspects as VV&A of those models. His belief is that a sound VV&A methodology will help to ask the right questions, to organise and structure the development and application to reach the intended goal. (e-mail communications with René Jacquart).

The REVVA methodology is designed to handle only model validity. Certainly REVVA is a good instrument to make a breakdown into goals and subgoals from the defined intended purpose of a simulation model, but REVVA is not referring to application validity which probably is the most interestingly when it is about HBRs and their uses.

There is a discussion going on in the REVVA group around the concept of validity. For example, some simulation models can be useful but unvalid, or valid but not useful etc. This is another way to express the different types of validity characteristics discussed in this report.

7 Conclusions and future work

This report reflects the current status of VV&A of HBR. It has drawn upon ideas and concepts from the field of experimental psychology and their more than 50 years of efforts to define different types and aspects of validity. There is a strong conviction among people within the HBR modelling community that validation of HBR does not always, if ever, requires the indistinguishability between human behaviour and the models behaviour. It is usually more of an interest to find the useful portions and functions to achieve the intended purpose.

Recently, a number of researchers have suggested that analyzing the goals or application of the model should serve as a starting point for defining appropriate validation activities (“application validity”; Campbell & Bolton in [Gluck & Pew 05]).

In many references the authors’ give their idea of of the HBR development - and validation - roadmap e g [Pew & Mavor] and [Gluck & Pew 05]. [NATO SAS017, 2001] gave this picture six years ago:

5.3.5.2 Validation Roadmap

Research on validation of human behaviour models in the short term (2000-2005) should lead to systematisation of current best practices. A number of efforts could help to address this point:

- Establish a process to develop testable requirements by using multi-disciplinary teams
- Development of an improved method for including SMEs in the validation process
- Define requirements of validation information in the catalogue that comes with the HBR
- Identify lessons learned from previous point-experiences in validation of human behaviour models
- Evolve procedures for validation (e.g. from Recommended Practices Guide – DMSO)
- Identify code of best practice

In the medium term (2005-2010) the validation area will be sufficiently mature as to develop informal methods and tools to support validation. This asks for:

- Development of a formalism to express the conceptual models for human behaviour representation
- Development of a formalism to express requirements for HBR
- Development of a toolset to support validation processes (e.g. automatic test case generation, knowledge-base integrity testing (consistency, completeness, coherence))

In the long term (2010-2015), on the basis of the foregoing developments, finally formal methods and tools to support validation may result. This implies:

- The development of formal techniques for validating HBR
- Extending and refining informal methods and tools to support validation

It is easy to get the impression that we still live in the short term task area. Validation is one of the four strategic and most important processes according to [NATO SAS017, 2001] besides knowledge acquisition on one axis and composability and interoperability on the other. It is formulated as:

There are two aspects of model development common to all application areas – knowledge acquisition and validation. At present the practice of acquiring knowledge upon which to build human behaviour models is time- and skill-intensive, resulting in incomplete representations. Similarly the validation of human behaviour models is time- and skill-intensive, often short-changed in the desire to complete the development cycle. The lack of useful tools and technologies hampers progress in knowledge acquisition and validation.

To enable more cost-effective development and to promote reusability across efforts, a more structured approach to human behaviour modelling is advocated. The human behaviour community must look to other research communities for tools and technologies that can be applied to human behaviour representation. The concept of model composability is cited as a key to cost effectiveness within mainstream modelling, but the concept has not been fully explored in the context of human behaviour modelling. Also the development of the High Level Architecture has allowed the interconnection of a diversity of simulations, but interoperable models of human behaviour representation have not yet emerged.

[Gluck & Pew 05] conclude their thoughts:
"… that, at the current state of development, resources need to be allocated not just to building the models, but also to a) collecting the knowledge and human performance data needed to make them function realistically; b) iteratively conducting formative and summative evaluations to ensure robustness, usefulness, and validity; and c) continuing to support new science leading to breakthroughs in concepts for improved architectures and more robust models. …it is short-sighted to support only the specific development of the models to the exclusion of supporting the research, quantitative validation studies, and infrastructure needed to improve the sophistication and scope of behaviours that can be represented in high-quality models. …considerable additional research is needed to achieve the desired levels of robustness, integrative fidelity, validity, parsimony, inspectability, interpretability, and cost-effectiveness. These research directions are more than worthwhile. They are imperative."

Acronymes

AI	Artificial Intelligence
CGF	Computer Generated Forces
EF	Experimental Frame
FEDEP	Federation Development and Execution Process
FOI	Swedish Defence Research Agency
FMV	Swedish Defence Material Administration
GOMS	Goals, Operators, Methods and Selection Rules
HBR	Human Behaviour Representation
HTA	Hierarchical Task Analysis
HF	Human Factors
IO	Information Operations
KBS	Knowledge Based System
M&S	Modeling and Simulation
MoE	Measure of Effectiveness
PMF	Performance Moderator Function
SISO	Simulation Standards Organisation
SME	Subject Matter Expert
SoI	System of Interest
UML	Unified Modeling Language
VV&A	Verification, Validation and Acceptance (or Accreditation)
VVE&T	Verification, Validation, Evaluation and Testing
V&V	Verification and Validation
WEAG	Western European Armament Group

References

[Brade 2004] *Generalized Process for the Verification and Validation of Models and Simulation Results*, 2004

[Cronbach & Meehl,1955] CONSTRUCT VALIDITY IN PSYCHOLOGICAL TESTS, Lee J. Cronbach and Paul E. Meehl (1955), First published in *Psychological Bulletin, 52*, 281-302. Internet available as Classics in the History of Psychology, *An internet resource developed by Christopher D. Green*, *York University, Toronto, Ontario at http://psychclassics.yorku.ca/Cronbach/construct.htm*

[DMSO, 00] Defense Modeling and Simulation Office. *Verification, Validation and Accreditation (VV&A) Recommended Practices Guide (RPG)*, 2000, http://vva.dmso.mil/.

[FEDEP] *Recommended Practice for High Level Architecture (HLA) Federation Development and Execution Process* (FEDEP), IEEE 1516.3.

[Gluck & Pew 05] Kevin A. Gluck, Richard W. Pew, *Modeling Human Behavior with Integrated Cognitive Architectures*, 2005.

[Gonzalez & Murillo] Gonzalez A., Murillo M., *Validation of Human Behavioral Models*

[Harmon et al] Harmon S.Y., Hoffman C.W.D., Gonzalez A.J., Knauf R., Barr V.B., *Validation of Human Behavior Representations.*

[Harmon & Youngblood] Harmon S., Youngblood S. *Validation of Human Behavior Representations*

[ITOP, 04] International Test Operations Procedure. *General Procedure for Modeling and Simulation Verification and Validation Information Exchange*. ITOP 1-1-002, WGE 7.2, 2004.

[Jacquart et al., 03] R. Jacquart, P. Bouc and D. Girardot. "A socio-technical view of a VV&A methodology", Euro SIW, 2003.

[James Denford 2005] *Experimentation Guide, rev 2*, AEC-N 0501, Kingston, Ont, Canada, 2005

[METHGU2, 04] *VV&A Methodological Guidelines Reference Manual*, THALES JP11.20 Technical Report JP1120-WE5100-D5101-METHGU2-V1.2, 2004.

[Mojtahed et al., 05] V. Mojtahed, M.G. Lozano, P. Svan, B. Andersson, V. Kabilan, *DCMF – Defence Conceptual Modelling Framework*, FOI-R--1754--SE, 2005.

[NATO, 98] North Atlantic Treaty Organization. *NATO Modelling and Simulation Master Plan, Document AC/323 (SGMS)D/2, Version 1.0*, 1998.

[NATO RTO SAS17 01] RTO *Technical Report 47 Human Behaviour Representation*, 2001.

[PDG, 06] VV&A PDG Working Group of the SISO Standards Activity Committee, *Draft*

Recommended Practice for VV&A of a Federation, an Overlay to the High Level Architecture Federation Development and Execution Process, 2006.

[Pew & Mavor 98] Richard W. Pew, Anne S. Mavor *Modeling Human and Organizational Behavior*, 1998.

[PROSPEC, 04] *VV&A Process Specification, User's Manual (PROSPEC)*, THALES JP11.20 Report JP1120-WE5200-PROSPEC-D5201-V1.3, 2004.

[Russell & Campbell et al] Russell S., Dorsey D., Ford M., Campbell G., Van Buskirk W., McCreary A., *Model Validation Is not Simple, Even When the Model Is: Lessons Learned From a Computational Model of Performance*, BRIMS 06

[Shannon, 75] R.E. Shannon. *Systems Simulation and the Art of Science*, Prentice Hall, Eaglewood Cliffs, N.J., 1975.

[Sullivan & Chew, 05] C. L. Sullivan and J. Chew. "*Documenting Verification and Validation Evidence with the International Test Operations Procedure on V&V*", Spring SIW, 2005.

[Yi et al., 06] C.-h. Yi, V. Mojtahed, M.G. Lozano, *REVVA and DCMF: A link between VV&A and Conceptual Modelling*, FOI-R--2110--SE, 2006.

[Ziegler et al., 00] Bernard P. Ziegler, Herbert Praehofer, Tag Gon Kim. *Theory of Modeling an Simulation*. Second edition. 2000.