

Session B3
**V&V (especially validation) for M&S with Human Behavior Representation
or People in the Loop**

Session B3 leaders:

Co-Chairs: **John Tyler** (MITRE) and **Sue Numrick** (DMSO)

Session Recorder: **Randy Saunders** (JHU/APL)

B3 Materials in Foundations '02 proceedings:

Papers

Introduction to Session B3: V&V for M&S with Human Behavior Representation or People in the Loop (1 page) [B3_intro]

Scott Harmon (Zetetix)

Archie E. Dillard (FAA Flight Standards Service)

Validation of Human Behavior Representations (34 pp) [B3_harmon]

S.Y. Harmon (Zetetix)

C.W.D. Hoffman (U.S. Army White Sands Missile Range)

A.J. Gonzalez (University of Central Florida)

R. Knauf (Technische Universität Ilmenau, Ilmenau, Germany)

V.B. Barr (Hofstra University)

Validation of Advance Flight Simulators for Human-Factors Operational Evaluation and Training (87 pp) [B3_dillard]

Archie E. Dillard (FAA Flight Standards Service)

Slides (may contain back-up materials and notes)

Validation of Human Behavior Representations (51 slides) [B3B_harmon in both pdf and ppt formats]

S.Y. Harmon (Zetetix)

C.W.D. Hoffman (U.S. Army White Sands Missile Range)

A.J. Gonzalez (University of Central Florida)

R. Knauf (Technische Universität Ilmenau, Ilmenau, Germany)

V.B. Barr (Hofstra University)

Validation of Advance Flight Simulators for Human-Factors Operational Evaluation and Training Programs (33 slides) [B3B_dillard in both pdf and ppt formats]

Archie E. Dillard (FAA Flight Standards Service)

Participants in this session are listed at the end of the Discussion Synopsis.

Discussion Synopsis (to provide perspective on papers & briefings identified above).

This synopsis examines points made by the authors as they presented their papers (with some commentary) and then provides more general commentary on the topic. Much of the discussion is in an outline format.

Paper 1

“Validation of Advanced Flight Simulators for Operational Evaluation and Training Programs”

Archie Dillard

- use airplane data to ascertain human performance, provides the inputs to the aircraft system model
- main discussion on validating process for flight simulators
- full six DOF motion system, visualization of environment for cockpit view displays
- FAA specified standards of performance (AC 120-40)
- Aircraft database representing flight characteristics, all subsystems
- Standards for data capture (does this include fidelity/accuracy guidelines?)

- Various types of simulations for flight ops

- History of Flight Simulators:
 - during WWII first significant use
 - transition to civilian
 - operations of simulator cost 1/10th of actual aircraft use
 - developments in computers, vis-sim, motion systems & databases,
 - Advanced Simulator Program (1970s)
 - Use of actual flight performance data, real-time capture

- Advanced Training Program
 - Level A through D (D is most capable)
 - D: requires no aircraft flight time for transition training
 - Aircraft and systems represented to highest level of fidelity possible – No Effort to Model Pilot

- Simulator Costs
 - Newest ones, \$14M for level D device
 - Includes: spares, training, tools and test gear, instructor/operator facilities, HLA/DIS compatible,

- Issues of use
 - Various flight performance issues in critical situations (wind shear)
 - Systems fidelity
 - Realistic environmental rep (weather, icing, cross-winds, etc)
 - Realistic faults/failures
 - Realistic operating environment
 - Realistic pilot workload

- Depending on the airline, they take different approaches and different application of procedures to define the operating environment that pilots must operate within. Consider new rules imposed after 9/11?
- Q: How do you define “realistic” and “highest degree of fidelity possible”? - to be addressed.
- Primary drivers?
 - new equipment certification rules
 - new airport designs (such as new different lighting techniques)
 - various anomaly conditions due to weather and other hazards
- Pilots as test subjects:
 - They have many requirements to meet and constantly must update their certifications
 - All pilots are qualified and the data they generate is within the qualification parameters
 - Highly trained, retrained
 - Very select and self-regulating population (elite qualifications)
 - But some Different airlines have different distributions of pilot age, other characteristics (ethnicity, cultural background, etc)
 - There are issues with training against certain anomaly conditions, because the airlines may not trust the results

Validation Test Sample:

- Aircraft manufacturer provides very detailed data sets for aircraft performance
- Testing of simulator is done against this data set
- If the simulator does not produce very similar data, within tolerances specified, then the test fails
- If any portion of the performance is out of tolerance range, test is failed

Validation of Navigation performance

- testing is done against geography around airports (terminal area geo database)
- on commercial airlines you must confirm geo-position along route
- Q: is the characteristics of weather at each terminal area used in testing? – YES.

System modeling:

- use manufacturer’s design data
- emulate aspects of system, using actual instrument display software
- Q: Hardware in the loop use? YES: use some aircraft hardware/instrument, eg flap indicators, (cheaper); also due to proprietary software (data is not available, but software system components can be used “as is”)
- Q: Can these trainers handle conditions such as carrier landings? YES.

Distributed Interactive Simulation (DIS)

- most simulators generally not HLA compliant, due to extra cost
- in next 6-8 years HLA will probably become fairly common and standard

- Simulators do not adhere to a standard or spec, they must meet the certification requirements at the end of testing
- Simulators are VERY price competitive – so options are rarely added

Operational Evaluation Programs

- low visibility ops
- new airports

Analysis of Results – examine all critical data

Advantages of Advanced simulators

- considered a high fidelity operational environment
- repeatable
- low risk to equipment
- low risk to human trainee
- cheaper than actual aircraft (generally, factor of 10 cheaper)
- better control of test environment
- equipment availability
- Q: How is radio reliability represented in FS? radio quality of reception is not simulated, the data is ‘canned’ so that it represent actual known flight radio conditions

Disadvantages

- still costly (\$300 to 1200/hr)
- due to demand, limited availability (though more available than aircraft)
- requires expert technical support
-

Examples of recent program

- Laser visual interference
- Worked with Brooks AFB labs, FDA
- FDA is responsible for regulating lasers
- Laser industry members
- Experimented with actual pilots
 - Three levels of exposure
 - What level is expected (5 uw is enough to produce flash blindness)
 - FDA specifies 2.5 mw as a level for adverse health effects)
 - Used 5, 15 and 50 uw as different levels
 - At 15 uw, partial flash blindness
 - At 50 uw, full blindness effect (temporary)

New Technologies on the Flight Deck

- heads up displays
- cockpit display of traffic info
- multi-function displays
- ADS-B
- Fly by wire systems

- Data link comms.
- Hazard avoidance/detection displays (audio)
- Navigation (GPS, LAAS, WAAS, - Local-area/Wide-area augmentation system)
- Communications
- Fly by wire technologies
-

Airport Design issues

- New Denver airport, many innovations
- Approach lighting changes
- High-speed taxi and exits (to reduce times on taxi way) – ongoing changes and approach testing
- Markings
- Land and Hold Short (safety issues are dealt with in simulation environment before policy is put into effect, after the anticipated results are replicated in
- Runway incursions
- Q: are there cases where a policy was developed with simulator results that went into practice and was later rejected because the results obtained in simulations was not observed in real use?
 - In Headsup display – some airports wanted to use HUD instead of airport lighting to alleviate some airport
 - The use of simulators in this way remains a very conservative system, such that many reviews and lots of data are required.
 - Location of runways (separation distance of parallel runways)

Environmental phenomenon

- wake vortex
- icing
- unusual attitudes

Summary

- use of simulator is useful in many, many situations and for many purposes, not just training
- open, collaborative environment between safety regulators, manufacturers, trainers, and

From Human perspective

- how much human knowledge is needed to design the simulator accurately?
- How accurate is human behavior/performance captured?
 - E.g., latency between human response and aircraft system response

What use of data on simulator sickness?

- in 60's and 70's this was a problem, but not in newer systems (for most systems)
- this was caused by high latency effects (response of simulator)

Paper Discussion Summary

1. Simulator cost about \$14M. Airbus aircraft sims cost about \$1M more due to more advanced onboard automation (computers). This is closing as Boeing adds more sophisticated computers, and Airbus adopts more typical human interfaces.
2. Pilot workload is one of the most difficult things to measure in a simulator.
3. Measurement of realism is a significant driver in sim certification.
4. Changes in pilot population will be hard to quantify as the military is no longer the primary source of trainer pilots. The regional airlines are the new source of pilots, the result is younger pilots but less experience in unusual attitudes or recovery from problems. The greater agility of military aircraft gave those pilots experiences they could draw on in emergencies.
5. Airport special cases are included in the training program of airlines that use those airports. One issue of environmental special problems like cool air inflow into San Francisco is not always well modeled.
6. Emulation as a technique to reduce computer hardware cost has not been scientifically studied, but subjective inspections seem good. Hardware emulations (instruments primarily) have been noted to be less desirable.
7. Proprietary software can make it very difficult to use emulated techniques.
8. Realistic simulation of communications problems, such as environmental radio interference, is difficult and also hard to accurately measure impacts.
9. Technology, such as electronic flight books, might replace maps some day, but since some planes have good displays and some don't. The lowest common denominator is still paper maps. There is a magic number somewhere, as older aircraft drop out of the population, where economies of scale make it possible to have an FAA mandate to use them.
10. The latency between action and response is a key measurable fidelity parameter. Human reaction times are 150-200ms, so there are latency limits of 150ms used in testing simulators. Other interesting fidelity issues, such as learning transfer, are much more difficult to measure.
11. Simulator sickness, caused by vestibular disconnects, was an issue in commercial sims in the 80s, but it is almost never a problem in the present advanced simulator environment.
12. Many policy topics have been investigated in the simulators. Laser flare was covered in the paper. Alaska used HUD information in lieu of airport lighting. There aren't any known cases where a simulator based policy made it into the field undetected.

Validation of Human Behavior Representations

Scott Harmon
Dave Hoffman
Et al.

Myths about HBRs

- Humans are good sources of requirements for HBR
- A good referent for HBR performance is a human doing the same job
 - In fact, a lot of specs state: “perform as well as” a human doing this job
 - But HBR are abstractions, not the real thing, don’t compare an abstraction to the real thing
- A valid HBR is as realistic as possible
 - I.e. error = 0
 - For all dependencies, all dimensions
- A good HBR is stochastic, like humans
 - But in fact, human behavior is NOT stochastic, and you don’t want it to be stochastic
- A good HBR is logical, like humans
 - But in fact, humans do not behave in a logical fashion, don’t think according to formal logic
 - Humans can manifest logic, but not constrained by logic or consistently express logic
 - Humans can express rationale best in Story-Telling mode, humans are very good story tellers
- An HBR “Fair Fight” is a clear and testable criteria
- Expert will recognize a valid or invalid behavior when they see it
 - But people can recognize a “believable” behavior or a “realistic” behavior
- Validating HBR is always very expensive
- Validating HBR is too hard
 - This is true in the very large, but not in the very specific, and HBR to be of value must be very specific
-
- Neural networks are non-linear classifiers, do not operate on logic, and can produce very useful decisions
- Reasoning may be performed in many approaches, not solely by logic

HBRs are unique among complex systems

- very high inherent complexity (similar to real humans)
- highly interactive and interdependent with surrounding systems (environment, platforms, tools, etc)

Validation is Challenging because:

- interact with complex environment
- deal with very large behavioral hyper-spaces
- Inherently are non linear
- Use oblique model representations
- Couple effects with other systems for all outcomes

1. Very little (or no) theory on constructing HBRs in the literature
2. Tools and techniques, exist, but largely in infancy (note for trace execution, or for KA/KE)

How close is the simulation to the simuland? (Cat's mental model while observing a bird)

- the referent is often not the actual thing, but some perception of it
- this introduces uncertainty and probabilities of error along many dimensions

But all of the above, can still reduce the highly complex space to a basic model

HBR Canonical Model

- humans as information processing systems (or machines)
- HBR behavior engine performs:
 - Accept inputs
 - Execute decision function
 - Generate actions that are executed (change internal state and effect changes in external environment)

There are four basic components in the canonical model

1. Human sensors – very complex and sophisticated
2. Behavior Moderators
3. Decision engine
4. Action system

Behavior Moderators

- a condition that affects human behavior in ways other than by cognitive elements
- Internal moderators (intelligence, experience, aptitudes, ...)
- External moderators (physiological stressors, environmental conditions, ...)

HBR Behavior Engine

- knowledge base element execution (a decision process)
- emotional effects manifestation
- Performance limitations (e.g., reaction times,

HBR Validation process (not a sales presentation)

Requirements? Validation Criteria elaborate the required simulation capabilities.

- derived from a user's purpose, but not solely based on the user's purpose
- users are NOT the best place to get the validation criteria

Users don't need HBR!

- they need specific capabilities
- they must express the capabilities accurately, but have limited ability to do so

So,

- pressure the user
- invent the criteria, OR
- Derive from automation needs

So, an HBR is an automation problem and a system control problem, and we know a lot about these

Partition the HBR in 3 parts

- Human Roles (mission, job, task, action hierarchy)
- Cognitive Functions
- Non-Cognitive functions

Military Operations (roles) taxonomy

- Combat Functions
 - C4ISR
 - Combat Ops
 - ...
- Combat Support Functions
 - Engineering
 - Policing
 - Transport
 - ...
- Combat Service Support Functions
 - Maintenance & repair
 - Medical
 - ...
- Non-Combatant functions
 - ...

Non-cognitive dependencies

- Physical factors
 - Weapons effects
 - Weather effects
 - Sensory inputs
 - Human factors
 - ...
- Psychological Factors
 - Capabilities
 - Emotional response
 - Complex factors

Cognitive Functions

- Situation understanding
 - Measurement
 - Assessment

- Prediction
- ...
- Plan Construction
 - Complications
 - Basic planning
 - ...
- Plan Execution
 - Actions
 - Execution conditions

Lessons Learned from Practice

- poor HBR requirements specifications create a domino effect – barriers to achieving HBR validity
- validation decisions cannot be defended without effects and performance requirements that drove the implementation
- requirements specifications usually do not specify effects and performance measures
- some acquisitions purposely avoid specific requirements
- Lacking other specification, HBR KA attempts to describe its domain from a “reality” perspective, using expert parlance
- HBR KA is too general in description
- Development can never achieve reality
- Users and decision makers incorrectly assume that validity is related to reality

Referents:

- Define the standards to gauge accuracy
- SMEs will provide different answers at different times to the same question
- Q: is there no reliable expert testimony?
 - Barr Act (congress rules on use of expert testimony) is under revision
 - Legal use of these rules is contentious
 - Most court cases involve dueling experts, leaving the decision to the judge & jury
 - Who was more credible, believable?
- Six levels to organize available referents
 - Domain, sociological, psychological, physiological, computational,

SME & Experiment referents

- Pros & Cons for both of these these

Validation bifurcates into 2 domains

- a) validation of design (documented requirements, assumptions, conceptual model)
- b) validation of implementation (observable results, measurable outcomes)

Why validate the Conceptual Model?

- verification is only possible if you have one, not possible if you don't
 - if the CM is not validated, you don't really have a CM
- The information you get from a design review is Very Subjective!

KB Validation?

- the knowledge base makes an essential contribution to HBR validation
- techniques for KB V&V have been thoroughly considered, rich state of the art
- Development of expert systems and KBS have been heavily invested
- Theory exists
- Techniques are robust
- Tools are efficient and useful
- Many problem areas have used the Tools and Techniques
- Lots of experience with the KBS applications developed, inspire confidence

Results validation

- you want to be able to rationally sample data from human performance
 - but performance is nearly always non-linear (cannot assume linearity)
- but no user will accept results if you throw out the data from human performance
- two basic problems:
 - functional complexity
 - unreliable observation and repeatability of observations

Recommendation:

- collect data at several points in each scenario
- with a SME interpretation of behavior (independent Verifier)
- Systematic data analysis leads to rational HBR validation assessments

Discussion Summary

1. Human logic is a style of reasoning that is used to rationalize actions, usually after the fact, rather than a technique used in human brain hardware.
2. Requirements taxonomy was built from documented requirements, but there are not use cases that show how those capabilities would usefully employed.
3. SMEs can provide different results under different situations, because they provide an aggregate of all the inputs received. If you ask for a narrow checklist for measures of performance, the results are better because there is less room for aggregation.
4. Expert testimony, in the legal environment, has started to build definitions of “expertise”. The key issue is differentiating when an expert is testifying about knowledge that is generally accepted within a scientific community, versus fuzzy boundary areas where answers reflect pet theories or hobby horses.
5. Will there ever be a “universal” HBR? Perhaps, within a reasonable range of applications. However, the track record for future predictions isn’t very good.
6. Why can’t we use the techniques for measuring human “goodness” for measuring HBR “goodness”? The officer who’s assessing human goodness has a lot of things to consider that aren’t replicated in HBRs. The coupling between these features is strong, and it helps the officer selection process.

7. Why do we want a model of human behavior that replicates human behavior, complete with foibles? Wouldn't we prefer a model of ideal human behavior?
8. What if we need "geographically typical and demographically representative" in simple clutter or other applications that aren't as complex as a HBR representing a general? The approach must be to stick with the minimum requirements you must have. Adding additional requirements because you "might" need them, greatly changes the complexity of the HBR management and V&V problems, for little return.
9. All we should demand from an HBR is that it fits the needs of the intended use. Demanding it to be "the same" as people is unreasonable.
10. The HBR community might benefit from the use of "error bars" as an indication of when the behaviors are good enough.
11. Is "believability" a metric that we can hang our hat on, because in the entertainment industry doesn't require that the behaviors be most likely, simply that people think that what they are seeing "might" happen.
12. The model for humans inside "America's Army" is a real human. They have scaled this to almost 800,000 people - all of them believing the game is reasonable. The Army was the sponsor and the Moos Institute was the implementer.
13. The mission rehearsal and experimentation communities don't get along well, with MR focusing on "looks perfect" and experimentation is looking for "good enough".
14. V&V can be an offensive weapon as well as it's traditional role as a defensive weapon. The technique is to expect nay-sayers to prove that what they want to do "adds" to the quality of results.

Discussion Notes

Is it possible to build HBR to meet any particular need?

Yes- As long as the need is well-defined and the purpose of use is well and completely specified. However, there are some areas of need (some aspects of behavior representation) where we lack sufficient power and foundations to overcome inherent complexity

- So far the track record for specifying future systems needs is poor. Predicting needs of future users is very dicey
- Stick with the human behavior of current and past systems, because needs definitions are easier to establish
- Counterpoint: We do not have the ability to accurately characterise the problems of current HBRs, because terminology consensus is not sufficiently in place, and because lack of agreement on foundations (theory) is endemic.

Why do we want HBR to replicate human behavior? Is this really desired? Do we want to really represent human errors, foibles and frailties?

- Do we want to play chess the way people really play chess?
- You must understand the relationship between the process of coupling the HBR model/simulation with the actual
- Is the way to get HBR to first get real humans and measure in great detail their behavior?
 - YES: if you are modeling physical actions and interaction with environment
 - NO: if you are modeling cognition and non-cognitive behavior components of the simulation

There is a Need for simulation of “clutter” in military operations

- huge numbers of entities that confuse and generate non-linear dynamics, “fog of war” effects
- address political and social influences, other soft factors influencing decisions in ops planning
- e.g., what was the impact on an Israeli Cdr when Zinni was sent to Jerusalem?
- The question is “who you are stimulating with the simulation?”
 - IF a Decision Maker: THEN solution A
 - IF an operator: THEN solution B
 - So solutions are application and use-case specific
 - V&V must be use-case constrained
 - There is no common or general solution, or even a small solution set that satisfies the range of needs across the application domain
- The bottom line is the system being modeled is highly complex, orders of magnitude greater complexity than other systems (even advanced aircraft such as the USMC Osprey)
- Some assert that human behavior is more complex than behavior of atomic weapons

If the cost of developing the acceptable HBR is so great, due to this incredible complexity, then it is not cost-effective to develop HBRs, because it is more economical to employ real humans (in most cases) than it is to develop synthetic human behaviors

Again: Is the goal of HBR to provide replacements for Humans?

- not usually
- even with robotics, the application of robot is task-job specific (e.g., to perform recon in a hazardous space or to move into area where real human could not go and remain alive & healthy)

On the Need for HBR –

- we are caught in a relationship making comparisons between simulated results with operational results
- the operational results are almost always abstracted from actual results (after action reviews, battle damage assessments, lessons learned, etc)

Many historical examples of deviation from doctrine in combat. It is an axiom.

- Why can't we do in models what we can do in real life?
 - We are far more capable in our brains than we can express in codes in computer simulations
 - We cannot directly infer or inquire what the SME knows or how he/she decides

- From a practical point of view, if we have to get something done (build ONESAF) then we have to create a solution to satisfy the stated requirements
 - Schedule & cost pressure
 - Questions of representing human beings fall by the wayside
 - The only way to reduce the enormous problem, is boil it down to essential effects to produce, and focus on that capability
 -
- Validity has nothing to do with how right it is, it has to do with how well it suits the defined need
- So we have to start with clearly specifying the desired effect, stating the requirements
 - If we can meet these requirements, then the HBR is valid

VALIDITY = FIT FOR A STATED PURPOSE UNDER ANTICIPATED CONDITIONS WITH A KNOWN AUDIENCE/USER GROUP

We want to look for a “human effect” in the simulation. Without reliance on “stochastic” mechanism, are we looking for a fit to specific circumstances?

- are we close enough to be able to produce the desired effect?

The entertainment community has a different approach: focus on Believability, realism

- For the “America’s Army” game (MOVES Institute), developers were told to live under the conditions, and experience the actual environment they were to produce in the game
 - They did not rely on SMEs
 - The model of humans in the game, are actual characters of human players (HITL controllers in the loop over the net)
 - Long term goal is to have some “team members” that are synthetic
 - Success has been achieved thus far because the game quantifies the human behavioral inputs from human players into a limited set of (41) input parameters that can be continuously sampled and transmitted via internet at common modem speeds

The idea of a mission rehearsal is a huge stumbling block in simulation for experimentation

- they don’t go together very well
- experimentation needs approximation
- mission rehearsal needs very accurate representation of actual conditions
- and exacerbated by the “scab pickers” from the V&V practice community
- these are ex-post-facto efforts to define requirements to a degree beyond which satisfy the user community

We may find those who look at V&V as a defense, but we should view it as a weapon that can be wielded by the Sim developer agent. Challenge the V&V agent to identify some aspect of the requirements that are not met.

B3 Session Summary

The session included papers on two very different applications of humans in simulations:

- human in the loop simulators (HITLS) for airplane pilot training (and related FAA applications)
- HBR for synthesis of human in constructive simulation

Needs for V&V in these two domains are very different

Key Issues to motivate research:

Developing Data sets for HBR V&V

1. in HITLS, you can gather tons of data, very detailed, from the (aircraft) system that describes the system response to human inputs and reactions
 - few issues of reliability
 - much practical experience doing this (50+ years of practice, since the early 50's)
 - In Flight Simulator V&V
 - dependency on vast database of actual flight system operations,
 - great detail,
 - physics based modeling
 - in early flight system modeling, it was done with very detailed physical models, in wind tunnels and scale models
 - many years of practical experience, methods and tools before these models were put into computers
2. in HBR, there is very little data, and very limited ability to gather appropriate and reliable data
 - access to data, when it exists, can be very contentious
 - limited experience (20 years, maybe, between first CGFs and SAFs , from SIMNET to SCOTT
 - early AI efforts did not put great effort into validation, due to Turing Test
 - early models were based on human perception and judgement about behavior observed
 - theory is founded in psychology, where there are many competing theories explaining the same behaviors in very different terms
 - in current practice, the most well-validated human models are of physical attributes (DI-Guy, JACK, etc)
 - Successful use of a general data driven V&V model or process with HBR is beyond current capabilities
 - **High functional complexity makes data collection difficult**
 - **Directly observable behavior that provides very limited insight into internal functions makes data analysis difficult**
 - Research on these two fronts is needed

B3 Session Participants (17)

First Name	Last Name	Organization
Pierre	Bouc	ALEKTO
Anthony	Cerri	U.S. Joint Forces Command
William	Cox	Aviation Systems Concepts, Inc.
Archie	Dillard	Federal Aviation Administration (FAA)
Scott	Harmon	Zetetix
Camillus	Hoffman	TRADOC Analysis Center
Susan	Kirschenbaum	Naval Undersea Warfare Center
Andreas	Koester	ITIS e.V.
Eric	Lazur	JHU/APL
Sue	Numrich	DMSO
Larry	Sanders	DTRA
Randy	Saunders	JHU/APL
James (Chuck)	Segrest	Lockheed Martin
John	Tyler	The MITRE Corporation
Ted	Raitch	Anchor Trading International Corp.
Simon	Goerger	The MOVES Institute
Ron	Morishige	Lockheed martin