Appendix 1

Best Practices for Human-in-the-Loop Validation Exercises

Prepared By:

Andrew Harvey
EUROCONTROL Experimental Centre

Karen Buondonno, Parimal Kopardekar, Sherri Magyarits
Federal Aviation Administration William J. Hughes Technical Center

Nicole Racine
Titan Systems

3 February, 2003

Acknowledgments

The persons listed below attended the March 2002 workshop in Atlantic City, New Jersey. This paper is a collection of their ideas, experience, and expertise. The authors would like to thank them for their dedicated involvement and contributions to this effort.

BAE Systems - Virginia Saulnier
CENA - Didier Pavet
Deep Blue and University of Sienna - Patrizia Marti
EUROCONTROL Experimental Centre - Robin Deransy, Alistair Jackson, Barry Kirwan, Nigel Makins
ENAV - Giancarlo Ferrara
EUROCONTROL - Ulrich Borkenhagen, Hans Wagemans
FAA Civil Aerospace Medical Institute - Carol Manning
FAA Headquarters - Diana Liang, Jacqueline Rehmann, Diana Liang
FAA William J. Hughes Technical Center - Richard Ozmore, Ben Willems, Tanya Yuditsky
ISDEFE - Nicolas Suarez
MITRE CAASD - Celesta Ball, Urmila Hiremath
NASA Ames Research Center – Sandy Lozito
NASA Langley Research Center - Kara Latorella, Dave Wing
NLR - Michel de Bruin, Juergen Teutsch
San Jose State University – Lynne Martin
University of Sienna - Antonio Rizzo
Volpe Transportation Research Center – Kim Cardosi

Table of Contents

List of Tables and Figures

Executive Summary

The first Federal Aviation Administration (FAA)/EUROCONTROL Action Plan 5 workshop was conducted on March 19-21, 2002 to discuss and develop best practices related to human-in-the-loop (HITL) exercises.

The workshop was well attended by 14 European participants and 19 United States (U.S.) participants. The participants were all practitioners experienced in validation research and HITL exercises. The original intent of the workshop was to focus only on identifying best practices for real-time HITL simulations, however, during the course of the workshop, the practitioners agreed that many of the identified best practices could also apply to a variety of HITL exercises, not just real-time HITL simulation. The following topics were discussed:

1. Role of HITL exercises in the validation process,
2. General overview of HITL simulation process,
3. Managing HITL validation exercises,
4. Scope and fidelity considerations,
5. Experimental design considerations,
6. Airspace and scenario characteristics,
7. Subjective and objective data collection,
8. Sources of error and variance, and
9. Statistical and operational significance of results.

Each topic was introduced by a short briefing and followed by a discussion session facilitated by a moderator. The participating practitioners performed all briefings and moderation. After the initial discussions on each topic, the participants were divided into working groups to further discuss and identify elements of the best practices for the topic assigned to them. Their consolidated recommendations were presented to all participants for final discussion and consensus. All discussions were recorded.

The recommended best practices described in this appendix are the results of collective input from all of the European and U.S. participants. The intention is that these best practices serve as supplemental guidelines for experienced practitioners who perform HITL validation activities.

Though different from the order which they were discussed in the workshop, the best practices are presented in harmony with the five steps outlined in the High Level Methodology Approach defined previously in the main body of this document. The best practices from each of the topics are organized by the periods before, during, and after an HITL validation activity.

Overall, the participants felt that the workshop was very valuable and useful for their research endeavors. They also suggested that the further workshops should be arranged to discuss additional topics such as statistical analysis, metrics, fast-time simulations, modeling, safety assessments, and advanced concepts.

1 Background

The Federal Aviation Administration (FAA) and EUROCONTROL agreed upon an Action Plan 5 (AP5) in November 1997. The goal of AP5 is to determine a strategy for validating and verifying the performance, reliability, and safety of Air Traffic Management (ATM) systems. Both the FAA and EUROCONTROL had developed separate validation strategies. Recently the organizations have worked to blend the documents into a harmonized strategy plan for validation and verification that can be adopted universally and serve as a best practices guidance to research practitioners (see document that precedes this appendix). By adopting such best practices, the validation process can be improved while being accomplished more timely and more cost-effectively. The application of the recommended strategies and practices may sometimes require unique but minor aberrations based on cultural, fiscal, and organizational needs.

1.1 Purpose of the Appendix

In addition to the overall strategy guideline, one of the objectives of the FAA/EUROCONTROL AP5 was to develop detailed best practices for all aspects of the validation process. Validation exercises take many forms including: analytic studies, concept studies, fast-time simulations, part-task and full-mission real-time human-in-the-loop (HITL) simulations, field-testing, and shadow-mode testing. Development and subsequent use of these best practices will allow for sharing of information and comparison of results. Best practices covering the use of metrics, data collection, data analysis, and reporting are needed for each type of validation exercise. These best practices take into account several factors including resources, cost, and most importantly, prior lessons learned. In order to capture lessons learned, a series of workshops will be organized. The attendees of these workshops will be experienced practitioners in their respective fields. The purpose of this document is to describe the best practices related to HITL exercises for validation of ATM initiatives.

1.2 First United States/Europe Workshop on Best Practices for Human-in-the-Loop Exercises

The FAA and EUROCONTROL organized the first AP5 workshop in March 2002. The objective of this first workshop was to discuss and identify best practices from practitioner experience with HITL exercises. There were 19 practitioners from the United States (U.S.) and 14 practitioners from Europe. The U.S. participants included researchers from the FAA William J. Hughes Technical Center, FAA Civil Aerospace Medical Institute (CAMI), National Aeronautics and Space Administration (NASA) Ames Research Center, NASA Langley Research Center, Volpe Transportation Research Center, and MITRE. The participants from Europe included researcher practitioners from EUROCONTROL, EUROCONTROL Experimental Centre, Deep Blue, University of Sienna, Ente Nazionale di Assistenza al Volo (ENAV), Centre d'Etudes de la Navigation Aérienne (CENA), Ingeniería de Sistemas para la Defensa de España (ISDEFE), and Nationaal Lucht en Ruimtevaartlaboratorium (NLR). The results of the workshop are documented in this appendix.

2 Air Traffic Management Initiatives and Validation

Sponsors, managers, and researchers often face the difficulty of determining needed validation exercises for an ATM modernization program. The scope and level of validation vary according to the type of the program. Generally, the modernization program that aims to achieve a positive change can be categorized as:

- Development of decision support tools (e.g., Traffic Management Advisor, Medium Term Conflict Detection),
- Procedural changes (e.g., Reduced Vertical Separation Minima),
- Advanced concepts (e.g., Dynamic Resectorization),
- New software/hardware (e.g., Standard Terminal Automation Replacement System, Display System Replacement, Medium Term Conflict Detection), and
- Advanced technology (e.g., Global Positioning System, Data Link).

2.1 Impact on the Human Operator

Validation activities vary for different types of programs. The level and importance of validation efforts also vary. They depend upon the potential changes that a modernization initiative may produce. The impact of an ATM initiative on the human operator is of primary importance and is therefore a key element of the validation process. Table 1 describes a sample of HITL impact factors potentially generated by a modernization program.

Table 1: Human-in-the-Loop Factors

The following is a list of factors that could be impacted by an ATM modernization program (FAA Human Factors Job Aid, 1999):

1. **Workload:** Operator and maintainer task performance and workload.
2. **Training:** Minimized need for operator and maintainer training.
3. **Functional Design:** Equipment design for simplicity, consistency with the desired human-system interface functions, and compatibility with the expected operation and maintenance concepts.
4. **CHI/HMI:** Standardization of computer-human interface (CHI)/human-machine interface (HMI) to address common functions, employ similar user dialogues, interfaces, and procedures.
5. **Staffing:** Accommodation of constraints and opportunities on staffing levels and organizational structures.
6. **Safety and Health:** Prevention of operator and maintainer exposure to safety and health hazards.
7. **Special Skills and Tools:** Considerations to minimize the need for special or unique operator or maintainer skills, abilities, tools, or characteristics.
8. **Work Space:** Adequacy of work space for personnel and their tools and equipment, and sufficient space for the movements and actions they perform during operational and maintenance tasks under normal, adverse, and emergency conditions.

9. **Displays and Controls:** Design of displays and controls (to be consistent with the operator's and maintainer's natural sequence of operational actions).
10. **Information Requirements:** Availability of information needed by the operator and maintainer for a specific task when it is needed and in the appropriate sequence.
11. **Display Presentation:** Ability of labels, symbols, colors, terms, acronyms, abbreviations, formats, and data fields to be consistent across the display sets, and enhance operator and maintainer performance.
12. **Visual/Aural Alerts:** Design of visual and auditory alerts (including error messages) to invoke the necessary operator and maintainer response.
13. **I/O Devices:** Capability of input and output devices and methods for performing the task quickly and accurately, especially critical tasks.
14. **Communications:** System design considerations to enhance required user communications and teamwork.
15. **Procedures:** Design of operation and maintenance procedures for simplicity and consistency with the desired human-system interface functions.
16. **Anthropometrics:** System design accommodation of personnel (e.g., from the 5th through 95th percentile levels of the human physical characteristics) represented in the user population.
17. **Documentation:** Preparation of user documentation and technical manuals (including any electronic HELP functions) in a suitable format of information presentation, at the appropriate reading level, and with the required degree of technical sophistication and clarity.
18. **Environment:** Accommodation of environmental factors (including extremes) to which it will be subjected and their effects on human-system performance.

2.2 Overview of Validation Techniques

Typically, the validation process involves multiple methods, techniques, and tools. The scope and resources needed vary depending on the level of maturity and type of ATM initiative that is being validated.

The following techniques are typically employed for validation exercises:

1. ***Concept studies/Paper studies/Analytical studies*** - Concept studies are performed at particularly early stages of an ATM initiative. These studies address technological feasibility, engineering analysis, benefits (or hypothesis), and analysis of a concept or an initiative. Paper studies are used for conducting a target level of safety analysis, risk analysis, cost-benefit trade-off analysis; and theoretical examination of aspects of the concept of operations. These studies may include gap analysis, functional decomposition, comparative studies, analytical studies, etc.
2. ***Task analysis*** - A task analysis focuses on identifying detailed tasks and subtasks. Such a detailed breakdown is useful for determining the division of activities within a team (e.g., executive controller and planner controller or Radar Controller and Radar-Associate). The breakdown is also useful for identifying any routine activities that are potential candidates for automation.

3. ***Storyboarding*** - The storyboarding technique is primarily used in early stages of a concept or ATM initiative. This technique is typically performed by a team consisting of subject matter experts (SMEs), human factors researchers, engineers, and other necessary members. This technique is particularly useful for examination of concepts and considerations such as information flow, procedural needs, and decision support tool functionality. Typically, the storyboarding involves drawing sketches and pictures (as if a story about a concept is being described with pictures, hence the name storyboarding) that are used for describing the concept or used as talking points to generate discussion about the concepts.
4. ***Cognitive Walkthroughs*** - Cognitive walkthroughs are also used in the early stages of concept exploration. Cognitive walkthroughs are used to discuss issues such as human error potential, data/information flow between operators and between operators and machines, information needs and decision support tool functionality, and procedural considerations. Storyboarding, task analysis, data flow diagrams are some of the specific techniques that are used during the cognitive walkthrough process. Typically, cognitive walkthroughs are performed by a team of SMEs, human factors researchers, engineers, and other members as appropriate. The walkthroughs provide a structure for team members to mentally imagine the concept and walk through the details of the ATM concept (hence the term cognitive walkthrough). The cognitive walkthroughs are useful during design reviews and identification of potential issues with integration of multiple decision support tools and procedures.
5. ***Fast-time simulation/modeling*** – This technique is based on human models but no human interaction is employed. All scenarios are compiled via computer-based simulation. Fast-time simulation and modeling exercises are typically performed to examine system performance including benefits assessment (e.g., delay, fuel burn, time/distance flown), and the analysis of capacity, safety, risk and efficiency. They are often used in the early stages of validation to get initial preliminary ideas of potential benefits. Fast-time and modeling studies are also useful for identifying potential problem areas where real-time simulation studies are necessary for further exploration.
6. ***Rapid Prototyping*** - Rapid prototyping studies provide an opportunity to develop HMIs for advanced concepts and conduct usability studies. Although rapid prototyping can be considered a development activity, the user-interface is often tied with the human and system performance**.** In some cases, rapid prototyping exercises will provide input to real-time HITL simulation studies and vice-versa. Prototyping is typically used to demonstrate the look and feel of an interface that will support a new technology or a concept. Increasingly, iterative prototypes are being used to identify requirements for user-interfaces. During the iterative prototyping process, SMEs are provided with initial prototypes of interfaces and based on their input the prototypes are modified. This process is used to generate team consensus on user-interface requirements.
7. ***Surveys and interviews*** - Surveys and/or interviews of SMEs are performed to gather their perspective on new initiatives. They focus on their opinions about feasibility, benefits, and acceptance. Such data can be useful, although not sufficient alone, to modify a concept.
8. ***Real-Time HITL Exercises***
   8.1. ***Part-task Real-Time HITL studies*** - Part-task real-time HITL studies are performed to examine a specific topic or question. For example, a part-task, real-time HITL

study may focus on issues particularly related to ground-side issues involving only air traffic controllers, or it may address air-side issues involving only flight crews. They are generally performed to identify and assess specific human performance issues as a result of new ATM initiatives (e.g., impact of new data link technology on communications duration).

8.2. ***Full-mission Real-Time HITL studies*** - Full-mission real-time HITL studies are performed to examine or demonstrate an end-to-end concept. For example, a full-mission real-time HITL study will likely include cockpit simulators, air traffic control simulators, flight crews, and air traffic controllers. The scope and full-mission varies based on the objectives of the particular study or demonstration. For example, a study may include simulation of entire day's ATM activities for a particular airport to assess capacity increases based on the addition of a new runway. A different kind of full-mission study may focus on a specific concept such as "shared separation" and its impact on the flight crew and air traffic controllers simulated in generic airspace.

8.3. ***Shadow-mode testing*** - Shadow-mode testing is particularly useful for new hardware and software initiatives. In this technique, an operational prototype is fed live data but is not used to control live events. This technique involves the simultaneous operation of both old and new systems in the operational setting. Such side-by-side testing allows for a 'real-world' evaluation of stability, reliability, performance, and acceptability of a new technology while the old technology is still operational, hence the term shadow-mode.

8.4. ***Operational Trials*** – In this technique and operational prototype is fed live data and is used to control live events. Operational trials are performed to demonstrate the feasibility and benefits of advanced concepts such as new technology, procedures, or decision support tools. These trials are usually performed in the later stages of the validation process. Initially, operational trials are performed under nominal scenarios such as good weather, low traffic volume, etc. to ensure the feasibility of the technology (e.g., data link, ADS-B). As the initial trials become successful, the operational trials are further performed under increasingly complex situations.

It should be noted that not all of the activities previously described are considered HITL techniques, however all of them to some extent rely on empirically derived human models (e.g. fast-time simulation).

Within an ATM modernization program, validation exercises should be carefully organized as part of an overall validation strategy plan. Figure 1 depicts a sample validation route map (commonly known as the Banana Model) that illustrates how various validation activities could be linked together. It must be noted that the different validation activities can be performed sequentially, in parallel, or iteratively depending on the need and scope of validation exercises, and you don't necessarily have to use all of them.

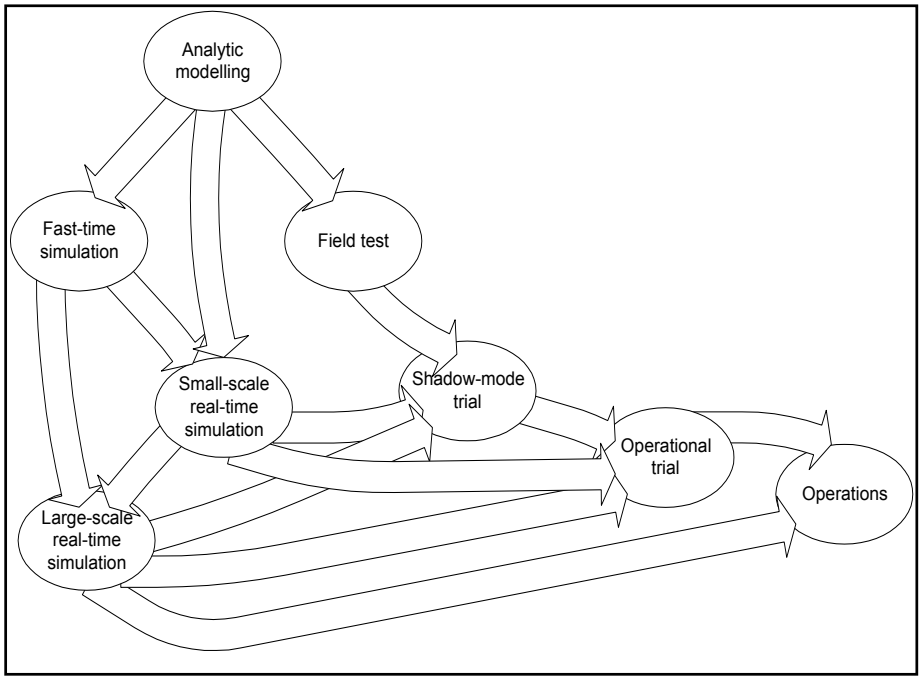

Figure 1. Sample Validation Route Map (Banana Model)

2.3 Choosing a Validation Technique

As previously stated, validation is an iterative and incremental activity. Advanced ATM concepts often require complex understanding of operations and their implications on procedures, decision support tools, and human factors. It is unlikely that one exercise will answer all the validation questions. Therefore, the roadmap to validation should comprise a series of validation activities, many of which will be involve HITL studies. It is recognized that the maturity of an advanced concept is one of the factors in deciding the type of technique that is suitable for the validation. There are a number of ways to describe the maturity of a concept. For example, NASA uses Technology Readiness Levels (TRLs). Another method is to use the five basic validation steps proposed by FAA/EUROCONTROL Action Plan 5 to describe the maturity of a concept. Table 2 summarizes the validation steps (labeled V1 through V5) and suggests different validation techniques that can be used at different phases of the process. It must be recognized that these steps are very generic and need to be interpreted according to the concepts and individual agency's processes. The validation steps are defined in more detail in Section 5.2 of the main body of this document.

Table 2. Validation Steps and Corresponding HITL Techniques

Validation Steps	HITL Technique
V1- Basic principles of new concept	Interviews, surveys, cognitive walkthroughs, data flow diagrams, task analysis, storyboarding, and analytical studies
V2- Initial proof of concept	Cognitive walkthroughs, functional decompositions, storyboarding, analytical studies, fast-time modeling, demonstrations, prototyping, and part-task HITL simulations
V3- Pre-operational demonstration	Part-task HITL simulations, full-mission simulations,
V4- Factory Acceptance	Full-mission simulations, shadow-mode testing, operational trials
V5- Onsite Validation	Shadow-mode testing, operational trials

2.4 Choosing a Simulator and Assessing Fidelity

Once the validation strategy and methodology are in place, and the appropriate technique is chosen, researchers are faced with the question of which simulator(s) should be selected (i.e., offers adequate fidelity) for their validation activity. This is an important consideration since it has implications on an activity's output, data precision, acceptability, and cost. In general, as concepts become more mature, the required fidelity of HITL exercises increases.

Simulator fidelity can be divided into functional and physical aspects. Functional fidelity refers to the functions and capabilities of a simulator as compared to their counterparts of the real-world operational system that is being simulated (e.g., a fuel-burn model of a cockpit simulator). The functional fidelity is very important for fast-time simulation studies. Physical fidelity refers to the appearance and human-machine interfaces of a simulator as compared to their counterparts of the real-world operational system that is being simulated. The physical fidelity is particularly important in HITL simulation studies. HITL simulation studies involve human participants interacting with the systems or simulators. Typically, HITL simulation studies collect data on human and system performance. Therefore, it is important that the "look and feel" of a simulator is very accurate and representative of the real system. Another aspect of HITL simulation studies is the participant fidelity. If the study participants do not accurately represent the study population the results may be biased. Participant fidelity is closely related to the statistical sampling principles, e.g. study participants should closely represent the experience, age, gender, and other important attributes of the target population.

The conventional method of fidelity assessment is to classify the fidelity of a simulator as low, medium, or high. This classification is loosely based on the presence or absence of

certain simulator attributes (e.g., avionics, range of motion, display capabilities, etc.). Another approach classifies cockpit simulator fidelity using defined certification levels (Level A, B, C, or D) based on characteristics and attributes. For example, a Boeing 747 cockpit simulator possessing the highest fidelity classification for its class of aircraft (Level D) is typically certified for airline training exercises.

The advantage of these methods is that they are very easy to understand and apply. However, these methods neglect the fact that not all studies require the highest fidelity in all attributes, and that the required fidelity of a simulator depends on the application under investigation. For example, consider two cockpit simulators. The first simulator has six degrees of freedom for motion and the second simulator has no degrees of freedom. However, both simulators have the same avionics and the same cockpit displays. These two simulators will certainly have different fidelity for a motion sickness assessment study but will have the same fidelity for a display layout assessment study. Clearly, assessment of fidelity depends on the attributes of a simulator that are useful to the simulation objectives. It is important to realize that even if a simulator perfectly represents a real-world operational attribute (e.g., six degrees of freedom for motion), if that attribute is not required for a specific simulation application, it does not contribute to fidelity.

Other methods exist to determine whether an available simulator offers adequate fidelity to conduct a particular activity. The steps below describe one method to determine the adequacy of a simulator.

1. *Identify the attributes that are important to the study objectives.* For example, if it is an air traffic control display simulator, it may be important to realistically represent the rate of aircraft turn, rate of climb and descent, aircraft data symbol, etc.
2. *Determine the importance of these attributes in a simulation on a 1-7 rating scale*. The importance rating can be received from users or subject matter experts. A rating of 1 on the scale means very low importance, 4 indicates moderate importance, and 7 indicates very high importance. The importance ratings may vary from one study to another depending on the study objectives.
3. *Determine the performance of these attributes of a simulator in a test on a 1-7 scale.* In order to judge the performance, a representative test must be conducted. This test will involve a study scenario. For example, an air traffic control display will involve display of aircraft operating in certain airspace. The performance rating can be received from users or subject matter experts.
4. *Draw an importance-performance matrix, where importance is in columns and performance is in rows.* The attributes are filled in the matrix with respect to their ratings. An example is shown below.

Table 3 provides an example of an importance-performance matrix for fidelity assessment.

Table 3. Example of an Importance-Performance Matrix

Performance Rating	Importance Rating						
	1 Very Low Importance	2	3	4 Moderate Importance	5	6	7 Very High Importance
1 - Very Low Performance							*Climb rates*
2							
3							
4 - Moderate Performance				*Turn rates*			
5							
6							
7 - Very High Performance							*Aircraft data symbols*

Table 3 indicates that this simulator has high performance and high importance for the presentation of aircraft data symbols. It has low performance but very high importance for the representation of aircraft climb rates. The latter observance indicates that the simulator in question is not adequate for the study. Typically, a simulator will be adequate if all of the important attributes (4 or above rating on the importance scale) have good performance (4 or above rating on the performance scale). If high importance is desired but low performance is experienced (3 or below rating), the simulator is not adequate for the application.

A mathematical method based on normalization of attribute values can also be used to assess the fidelity of different simulators. The method will select a simulator based on the highest fidelity needed. This method is computationally intensive and laborious but produces a numeric assessment of fidelity. The method can be found in Kopardekar 1999.

3 Best Practices for HITL Exercises

Although the sessions of the workshop from which these best practices have been assembled were organized differently, it will be more useful to the practitioner to regroup the practices according a commonly understood validation methodology. Therefore, the following sections present best practices for HITL exercises in harmony with the five steps outlined in the High Level Methodology Approach defined previously in the main body of this document (Section 5.2 or the main document). The best practices are organized by the periods before (corresponding to steps 1 and 2), during (corresponding to step 3), and after a validation activity (corresponding to steps 4 and 5). In addition, managing HITL exercises can present many challenges throughout the process, so there is a management element included in each of these sections.

The intention of the authors is that these best practices serve as supplemental guidelines for experienced practitioners who perform HITL validation activities. It is also recognized that this paper may not contain an exhaustive list of all best practices associated with validation research, and that some of the best practices listed could benefit from further refinement. The original intent of the workshop was to focus only on identifying best practices for real-time HITL simulations, however, during the course of the workshop, the practitioners agreed that many of these best practices could also apply to a variety of HITL exercises, not just real-time HITL simulation.

4 Best Practices for Before the HITL Exercise

Table 4 presents a list of recommended best practices to be considered when identifying requirements, and planning and preparing for an HITL validation exercise. Each best practice (labeled as *bp1*, *bp2,* etc.) is described in more detail following the table.

Table 4. Best Practices for Before the HITL Exercise

4.1	**Managing the Process**
bp 1	Identify the stakeholders, define their roles and responsibilities, and ensure good communication between them.
bp 2	Generate, compare, and prioritize lists of questions and/or concerns for each category of stakeholders.
bp 3	Map stakeholder questions into study requirements, then into simulation measures.
bp 4	Meet with controllers and pilots to discuss preliminary 'concept of use' issues before the HITL exercise, if feasible.
bp 5	Develop back-up plans for test design issues in the event of system problems/failures.
4.2	**Experimental Design Considerations**
bp 6	Make a clear statement of the type of exercise that is being performed.
bp 7	Produce a scientific justification for the use of the chosen HITL technique.
bp 8	Produce a statement of any constraints that apply to the design of the experiment.
bp 9	Define detailed and unambiguous objectives, and state hypotheses.
bp 10	Keep the number of objectives to a minimum.
bp 11	Document those areas where the exercise needs high fidelity to the real world.
bp 12	Make appropriate use of baselines.
bp 13	Be aware of the Human Factors issues of the design.
bp 14	Identify all sources of error and indicate those that can and should be controlled.
4.3	**Airspace and Scenario Development Considerations**
bp 15	Be open-minded. Consider unconventional options such as “generic airspace”.
bp 16	Research questions should drive scenario development.
bp 17	Identify and maintain common scenario characteristics for comparison.
bp 18	Starting and/or ending scenarios slowly is not usually efficient.
bp 19	Define and maintain necessary levels of realism.
bp 20	Traffic peaks and troughs may be relevant to research.

4.1 Managing the Process

The practitioner, particularly a Principle Investigator, has many responsibilities regarding managing the process of an HITL exercise. These responsibilities are essential to the success of an exercise and start early in the planning process.

bp 1. Identify the stakeholders, their roles and responsibilities, and ensure good communication between them

One of the major challenges involved with managing HITL activities is dealing with multiple stakeholders. A number of stakeholders can be involved in the planning, conduct, and analysis of HITL simulations, including the following:

- Investors/Sponsors
- Managers (e.g., air traffic control centers, regulatory agencies)
- Controllers
- Pilots
- Union representatives
- Suppliers
- Experimenters

Often a core exercise working group with representatives from each stakeholder group (or subset) is established. They actually "do the work" and facilitate the success of the exercise. The members of such a core team should be consistent throughout the duration of an exercise.

It is also of paramount importance that the roles and responsibilities are clearly defined for and mutually understood by all stakeholders and core team members at the onset of the activity. When problems occur in the HITL process, they can often be attributed to miscommunication between some of the stakeholders involved. Communication at and between all levels of stakeholders is key to meeting simulation objectives and achieving a common understanding of simulation outcomes.

bp 2. Generate, compare, and prioritize lists of questions and/or concerns for each category of stakeholders

Each of the stakeholders has a range of concerns for an HITL exercise, some of which overlap with other stakeholders and some of which do not. Generate a list of all stakeholder questions to ensure that, for example, the appropriate simulation design is constructed, or that certain events are scripted into the test scenarios.

Below is a sample of some common stakeholder questions (note: 'it' refers to whatever the simulation is testing):

Management questions:

- Does it improve key performance areas (e.g., safety, capacity, delays, human resources issues, training costs, manpower issues, selection issues, and quality of service issues)?
- Is it acceptable to the participants (e.g. controllers, pilots)?

Bottom Line: Will it improve matters operationally? Does it add value?

Controller/Pilot questions:

- Will it work operationally?
- Will it fit in with our working methods?
- Will I still enjoy the job?
- Will it make the situation more or less safe?
- Can I recover if it fails?
- Is it a threat to my work or career?

Bottom Line: Is it acceptable and useful? Does it add value?

Experimenter questions:

- Which, of the parameters it aims to improve, have improved? Were any of the controlled parameters affected; and were any other variables affected?
- Is the system acceptable to controllers/pilots?
- Are the cognitive skills of controllers/pilots being changed in any subtle or fundamental way?
- Will it work in the real world?
- Are there any side effects?
- Is safety maintained, improved, or degraded?

Bottom Line: What are the measurable benefits? How can we refine the design and/or operational requirements to make it better?

The HITL practitioner needs to have a common understanding of how the various questions are related to each other, where they overlap, and where they fit into the validation life cycle process. For example, a management question such as 'Is it safe?' may cascade down to the controller/pilot level into 'Is it safe in all circumstances, including system failures?' This question may then cascade down to the experimenter level into 'did a loss of separation occur? Did human error occur? Was situation awareness affected? What were the effects? Did workload increase or decrease as a result?' With careful planning, all of these questions could potentially be addressed with the same set of data.

bp 3. Map stakeholder questions to study requirements and simulation measures

Prioritizing stakeholder questions will also assist the practitioner in planning and executing an HITL exercise. This way, if constraints occur in simulation, for example, the experimenters can make trade-off decisions to best maximize addressing the 'highest importance' management questions, controller questions, pilot questions, etc. Practitioners should ensure early on that stakeholder objectives and questions are addressed by adequate

measures collected in the study. Identification of specific data requirements and the mapping of such data to objectives should be outlined in an experiment plan, before a study begins.

bp 4. Meet with controllers and pilots to discuss preliminary 'concept of use' issues before the HITL exercise, if feasible

Since controllers and pilots are usually the end users of the concepts evaluated in HITL exercises, their input can be an invaluable resource at the front-end of the simulation process. This stage is where techniques such as interviews with SMEs, storyboarding, and cognitive walkthroughs can be most effective.

bp 5. Develop back-up plans for test design issues in the event of system problems/failures

The HITL practitioner needs to plan for the unexpected. This is particularly true with complex, larger-scale studies involving the linkage of multiple laboratories and tools, and the coordination of multiple research organizations and study participants. Having a plan in place before the HITL exercise about how to troubleshoot system problems or work around 'no-show' participants will save time when the event(s) occurs and will salvage critical data based on predetermined priorities. Such a back-up plan might involve a modified test design, alternative data collection sources, stand-by participants, and/or "buffer time" in the laboratory schedule.

4.2 Experimental Design Considerations

An individual exercise usually forms part of a research program. The experimenter needs to be clear about where their exercise fits into the wider program and should have access to all other related experimental results. When designing the exercise it must not be forgotten that it will often be run using a simulation and not on the real system, yet any conclusions drawn will be expected to apply to the real system. Ensure that the design allows this mapping of results to the real world in the areas under evaluation.

The first step in the design of any experiment should be to determine the type of experiment required. Thereafter the experimental design considerations will depend primarily on the purpose of the experimental work.

bp 6. Make a clear statement of the type of exercise that is being performed

Types of exercises can be loosely categorized as exploratory, inferential, or demonstrative (formative or formal has also been suggested).

- Exploratory work may include techniques such as pilot studies, rapid-prototyping activities or other exercises where the main objective is to show the feasibility of a potential method or approach. The appropriate use of pilot studies will prevent unexpected problems from occurring during simulation execution and analysis.

- Inferential studies aim to detect differences between different systems under test. They usually have strict data requirements designed to permit statistical analysis of the results and thus require a high level of experimental control.
- Demonstrative HITL simulations focus on representing the "look and feel" of the system but have few, if any data requirements. They are usually performed toward the end of the development lifecycle with the aim of confirming human involvement and commitment (e.g., user acceptability). .

Though demonstrative studies are commonly used throughout the research community, for scientific reasons there should be more extensive use of inferential studies, which have more data rigor and statistical power for validation activities.

bp 7. Produce a scientific justification for the use of the chosen HITL technique

A scientific justification should outline the questions that will be answered by the study and why it is felt that a particular HITL exercise is the most appropriate means to explore a hypothesis. This may be justified by first conducting cognitive walkthroughs or task analyses that identify the need for further exploration. This justification may identify further work that can be achieved using existing models or fast-time simulations and thus help reduce the scope of a real-time simulation. Where analytical models have already been conducted, these should be used to form hypotheses and predict the results of the planned simulations. A major difference between predicted results and experimental results might indicate a problem either with the model or with the simulation design.

bp 8. Produce a statement of any constraints applying to the design of the experiment

Having confirmed the need to undertake the HITL exercise, the next stage is to consider constraints that will be imposed on the design. A common constraint, for example, is the duration of the study, which is often limited by the availability and cost of participating controllers, pilots, laboratories, etc. Another constraint might be that the simulation pilots (also known as pseudo-pilots) were not actual pilots and therefore may have responded differently to control instructions than a certified pilot would have. It is worth considering each potential constraint and to what extent each can influence the study. Knowing limitations prior to the experiment will help shape the design and allow for planning of contingencies for capturing every essential aspect of the study.

Once stakeholder requirements, the type of exercise, the reason for using real-time HITL simulation, and the list of constraints are established, the design of the experiment begins. Most of the recommended practices apply particularly to inferential studies, but should also be considered in the context of other studies types.

bp 9. Define detailed and unambiguous objectives, and state hypotheses.

It is imperative to produce a statement of Specific, Measurable, Attainable, Results-Oriented, Time-based (SMART) objectives. Not only should they be unambiguous but they should also map to higher level program objectives. This consideration is crucial so that exercise

results actually contribute to the overall validation of the concept. Ambiguous objectives may lead an exercise astray and dilute the results. In addition, hypotheses should be stated to match expected results of the activity (which are often formulated by taking into account results of previous, related work). Occasionally, an experimenter may choose not to draw specific hypotheses. For example, this may be the case for an exercise designed to explore a new advanced concept that is not fully defined. Whether or not hypotheses are appropriate, the exercise objectives should always be clear.

bp 10. Keep the number of objectives to a minimum

It is important that stakeholders, sponsors, and experimenters understand the risks of trying to address too many objectives in a single experiment. As previously stated, it is essential to produce a statement of the exercise objectives in clear and unambiguous form. It is also important to understand that there is a strong link between the duration of the HITL experiment and the number of objectives that can be usefully addressed. If the duration of the study is limited or fixed, then fewer objectives will result in greater confidence of results for each objective. Having too many objectives can misdirect to the focus of a exercise, put the exercise at risk for not having enough information to adequately assess a hypothesis, increase data requirements, increase the length of a study, etc.

bp 11. Document those areas where the experiment needs high fidelity to the real world

Not every aspect of a simulation needs to be of the highest fidelity. Cost and other factors often limit the level of simulation fidelity. However, it is not necessary to achieve the highest fidelity in every respect in order to relate the results to the real world. Careful consideration should be taken to ensure the level of fidelity is appropriate for each major area of the study. (See also Appendix Section 2.4 on Choosing Simulators and Fidelity Assessment).

bp 12. Make appropriate use of baselines

Baseline scenarios are often used to provide comparisons to a current operational procedure or system. However care should be taken to ensure that the simulated baseline is a sufficient representation of the real-world situation. It is also important to keep the contrasts meaningful. If the future scenario is too removed from today's system (in time or concept of operations) then the comparison to a baseline scenario of current operations may not be valid or useful.

bp 13. Be aware of the Human Factors issues of the design

In designing the experiment, the experimenter should be aware of the impact on the human participants. An over-demanding program of runs will cause fatigue and certain invasive measures will affect performance. Ecological aspects can be used to make the experiment more natural, for example, using a simulated shift change to test for situation awareness rather than artificially stopping the traffic. Minor oversights of the design can affect results if

they are not handled properly, for example provision of notepaper or pens. Such oversights will usually be detected if a pilot study is run.

bp 14. Identify all sources of error and indicate those that can and should be controlled

The main reason for using experimental design is to be able to control the sources of experimental error. Therefore the design must first identify the sources of error that are expected to be influential. Not all variance should be considered as unwanted. For example, in reality there is variance in sector entry times. If this variance is removed (as it can be in simulation) then the controller may start to 'learn' the traffic, which will have considerable (negative) impact on the results.

It is not possible to provide an exhaustive list of sources of error and variance. The following list indicates examples of the more common ones:

Traffic Familiarity: Re-use of the same traffic sample several times will introduce bias because the controllers will anticipate traffic behavior. This can be overcome by introducing perturbations such as weather and delay but this will add variance in the traffic behavior. Sometimes it helps to change minor things that don't affect traffic behavior such call signs or destination airports.
Simulation Piloting: Pseudo-pilots are usually too compliant and their voice communications are generally clear and constant. The use of real pilots, scripts, and synthetic voice generators can improve the situation.
Platform stability / fidelity: Continual interruptions due to platform problems will have a strong effect on controller/pilot attitude. If technical problems are associated with a particular period of the study this should be compensated for in the analyses.
Controller/Pilot attitude: Participants in a simulation can greatly influence the results. If they are insufficiently trained on new procedures or have low confidence in a tool this will have a negative effect. Conversely, participants who have been closely involved in the development of a particular tool or concept may be overly positive/negative in their responses. Some of these individuals may be particularly useful to help plan or develop an exercise, but it is best not to use them as test subjects as they may introduce bias.
Learning effects: The ability to work with new tools and procedures will almost certainly improve with time. Unfortunately, there is rarely enough time to train participants completely for advanced/future scenarios. Therefore, it is likely that participant performance in an experiment will improve with time. This can be compensated for by the use of blocking.
Order of presentation: In addition to learning affects, the order in which the experimental units are presented can be influential. Randomization techniques will minimize this affect.
Inter-controller/pilot variability: Differences between controllers can be compensated for by repeated measure designs. Where possible the experimenter should try to ensure that the participants are representative of the general controller/pilot population.
Intra-controller/pilot variability: Variable controller/pilot performance exists in the real world. However the experimenter should be cognizant that variability may be greater for participants in real-time HITL simulation due the effects of hard-to-control factors such as traveling fatigue, unusual working hours, and extended periods away from home. Although

it is recognized that intra-participant variability is difficult to control, sometimes it can be reduced by methods such as training to asymptote (i.e., training until the participants reach a bench mark criteria). It is also helpful to ensure that the simulation environment is similar to the operational environment, and that good experimental design techniques, such as randomization, are used to balance the effects of this variability.

4.3 Airspace and Scenario Development Considerations

Scenarios are usually characterized by airspace development, complexity, traffic realism, flight plan routes, traffic mix, overflights, special use airspace, weather, etc. Practitioners are challenged with creating airspace and scenarios that model the real operational environment, while addressing specific research questions.

bp 15. **Be open-minded. Consider unconventional options such as "generic airspace".**

In many HITL studies, actual airspace is replicated and real traffic scenarios are modeled to produce a familiar and realistic environment for data collection. For many exercises, using site-specific/existing airspace and traffic may be an essential requirement, such as when assessing the impact of a proposed new runway on operations at a specific airport. However, using site-specific airspace may induce constraints on the sample of subjects who participate in the study. Using generic airspace may be an option for studies that utilize many participants from various facilities or for general studies that do not need to be applied to a specific airspace.

Some of the advantages of using generic airspace include the availability of a greater number of participants. Very often it is difficult to use a large number of participants from any one air traffic control facility, airline, etc. because of staffing issues or participant availability. Another advantage is the ease of participant training on generic airspace. For example, consider an exercise where a particular airspace is modeled but controllers will be brought in from various facilities. Participants who work in the facility where the airspace is located will inherently be more familiar with the airspace than those from outside facilities. Generic airspace is generally easy to learn and ensures that all controllers become trained on the airspace at the same time. This means that the level of experience is the same for all participants.

Generic airspace is also more flexible than existing airspace since the experimenter has the ability to insert controlled obstacles such as weather, special use airspace, terrain, etc, and can easily create particular letters of agreement and standard operating procedures in order to address specific research questions. This approach may be very beneficial for concepts or decision support tools in the early development phase since these studies explore concepts that currently may not exist. The use of an unfamiliar airspace may help lift the participant out of present day procedures and be able to immerse them in a future concept environment.

Generic airspace does have some limitations. There are difficulties with the initial creation of generic airspace. Scenarios can be more difficult to build, multi-center facilities are more difficult to develop, and time must be allocated for the development of airspace procedures.

It may be difficult to baseline generic airspace scenarios and to generalize data for specific airspace. In addition, researchers will have to obtain sponsor buy-in for the use of generic airspace since results may be not be directly applied to any one facility.

It is not recommended to use generic airspace when you are researching a specific procedure for a specific facility. However, if generic airspace satisfies the requirements of an exercise, be open-minded and consider the approach. Know the advantages and disadvantages, and be prepared to justify its use.

bp 16. Research questions should drive scenario development

While the practitioner should strive to develop scenarios that model the operational environment, it is also crucial to build scenarios to address the research questions and objectives. All perspectives (researcher, operational, and management) should be taken into consideration.

bp 17. Identify and maintain common scenario characteristics for comparison

One of the biggest challenges in scenario development is to create scenarios that are experimentally comparable, but that are different enough to present a "new" problem for the participants. The experimenters often have to design scenarios that are similar (in complexity) but not exactly same. The similar scenarios are essential part of the good experimental design principles as they help statistical comparisons. However, scenarios that are too similar could produce (often undesirable) learning effects. The learning effects could negate the effects of experimental conditions. Simply changing the aircraft call signs are not always adequate since controller memory recognizes air traffic patterns as well. Learning effects may be controlled somewhat be randomizing the order of experimental conditions.

There is a great need for a common and agreed upon complexity metric that practitioners can use as a way to evaluate scenarios. In this way, a particular scenario may be compared with one that may be very different in some characteristics, but has the same complexity. Using the same measure of complexity would give a standard from which to potentially compare results from various studies around the globe. Usually, similar scenarios can be created by keeping several key factors consistent such as the number of aircraft, the number of conflicts, conflict geometry, type of aircraft, callsigns, and the number and type of structured and unstructured routes. It is also best to consider the opinions of SMEs during the scenario shakedown process to ensure that scenarios are comparable but yet not the same.

bp 18. Starting and/or ending scenarios slowly is not usually efficient

A common way of building a scenario is to initiate traffic gradually into the problem, allowing the controllers to ease themselves into the scenario. When the main part of the problem is over, traffic is usually tapered off. For an hour long scenario, as much as 15 minutes on both ends of the scenario might be dedicated to "ramp up/down" time. Given the time and cost restraints to run a simulation, this is not the most effective or efficient use of

scenario time. A more efficient means of building a scenario is to begin with a normal traffic load. Depending on the study, practitioners may also end the scenario in the middle of a conflict or other problem.

bp 19. Define and maintain necessary levels of realism

While developing scenarios, researchers are usually interested in maintaining high realism related to aircraft mix, traffic density, sector geometry, routes, and procedures. Many times, the scenario development process starts from collecting actual operational flight plan data. The operational data provides the realism. The operational data with realistic traffic density and aircraft types works well for near-term initiatives (e.g., Reduced Vertical Separation Minima). However, for future concepts such as free maneuvering, operational data may need considerable modifications in order to satisfy the experimental objectives. In such situations, a number of variables (such as future traffic load, future aircraft mix, and technologies) may need to be modified. Such modifications to better model the future environment may not be realistic in the present day operations, however will represent the future environment.

Care must also be taken during the conduct of the simulation. For example, if a scenario is developed such that it begins with a normal or heavy traffic load, practitioners may consider utilizing subject matter experts to brief a participant controller onto position. Since this is a normal operation for controllers, this would be an acceptable way to start the run with a full traffic load.

bp 20. Traffic peaks and troughs may be relevant to research

HITL practitioners often try to impose high workload levels on simulation participants to stretch the limits of the participant's abilities, and to test the limits of new concepts and procedures on the air traffic system. To do this, they typically design traffic scenarios to represent peaks or high levels of traffic activity. Slower-manifesting problems are generally avoided since the practitioner wants to make the best use of valuable time. Prior simulation research has shown, however, that operational errors often occur in the beginning of troughs, or lower levels of traffic, immediately following very high levels of traffic activity (this results because of a temporarily perceived reduction of complexity and thus lower vigilance).

In order to better emulate the operational environment and to capture all possible conditions for human error, HITL practitioners should script a range of traffic activity into their scenarios; those which include both peak levels of traffic and troughs.

5 Best Practices for During the HITL Exercise

Table 5 presents a list of recommended best practices to be considered when carrying out the tasks of an HITL validation exercise. Each best practice is described in more detail following the table.

Table 5. Best Practices for During the HITL Exercise

5.1	**Managing the Process**
bp 21	Do not sacrifice simulation quality for the interest of time.
bp 22	Document unforeseen effects of test variables on system performance not captured by system measures.
5.2	**Experimental Design Considerations**
bp 23	Insure adequate training of participants.
bp 24	Be prepared to administer contingency plans if necessary.
bp 25	Investigate concomitant measures carefully.
5.3	**Subjective and Objective Data**
bp 26	Be clear as to what is meant by 'subjective data'.
bp 27	Use objective data to clarify subjective findings (and vice versa).
bp 28	Explain to the participants why their feedback is needed and how it will be used.
bp 29	Understand the factors that could influence subjective data.
bp 30	Minimize the impact of external factors on subjective data.
5.4	**Validation Data Storage**
bp 31	Use a central data repository such as the Validation Data Repository.

5.1 Managing the Process

bp 21. Do not sacrifice simulation quality for the interest of time

A small set of good data is better than a large set of bad data. Whatever the time constraints may be, the HITL practitioner must remember this guideline and plan accordingly.

bp 22. Document unforeseen effects of test variables on system performance not captured by system measures

Though they may not be related directly to study objectives, researchers must be cognizant of unforeseen or subtle effects of simulation variables. Practitioners should document and report all observations and results, expected or otherwise, so that that may further educate and assist future management decisions.

5.2 Experimental Design Considerations

bp 23. Insure adequate training of participants

The value of training participants in an HTIL activity is often underestimated. It is generally desirable that participants clearly understand the objectives and the design of the experiment before they actually participate. This is often achieved by briefing participants prior to the start of an activity.

Most importantly, they must have sufficient laboratory familiarization and be adequately trained on any new procedures, equipment, etc. to ensure the validity of the results. When schedules are condensed, often the first thing to be reduced is the amount of time allotted for participant training. Do not compromise this aspect of simulation.

bp 24. **Be prepared to administer contingency plans if necessary.**

As mentioned in *bp 5*, it is a fact of life, particularly with real-time HITL exercises that some data or recordings may be lost due to technical reasons. The exercise should have been designed to accommodate a certain number of these losses without too much detriment to the exercise. The entire research team and laboratory personnel should be aware of the contingency plans in advance and be prepared to execute them should they be necessary.

bp 25. **Investigate concomitant measures carefully.**

The choice of measures (dependent variables) and experimental factors (independent variables) will depend on the objectives, and in turn the objectives should depend on the existence of suitable variables. The number of concomitant measures should be kept to a minimum to avoid the risk of finding significant effects that contradict simply by chance. It is difficult to explain such observances, especially to the stakeholders. However, it is a delicate balance because when several concomitant measures all indicate the same effect, this adds to the confidence in the results.

5.3 Subjective and Objective Data Collection

bp 26. **Be clear what is meant by 'subjective data'**

In any experimental work it is important to be clear about the type of data that is being collected. The experimenter should consider first whether any experimental benefit is to be gained from distinguishing between subjective and objective data types. In many cases these terms are misused to express the difference between qualitative and quantitative data. In the particular case of HITL studies, it can be useful to distinguish between the objectively measured system variables and those variables derived from the subjective opinions of the participants. An alternative terminology could be to define *perception* measures (subjective) and *observed* measures (objective). This definition also helps to clarify that the subjectivity is that of the human participant and not that of the observer.

bp 27. **Use objective data to clarify subjective findings (and vice versa)**

Clearly differences between the perception of the participants and the observed system recordings are of great interest to the experimenter. For this reason both types of data should be collected. In evaluation studies objective data can be used to generate discussion with the participants after an experimental run. Similarly subjective data can be aggregated statistically to produce classifications, for example, defining a "high workload" traffic sample based on subjective ratings.

bp 28. Explain to the participants why their feedback is needed and how it will be used.

The extent to which the subjective results will be used should be explained clearly to the participants. If they feel they are being asked to "sign off" on a new system, their feedback will be more guarded (and less constructive) than if it is made clear that the development is still in the early stages.

bp 29. Understand the factors that could influence subjective data

The attraction of subjective data is clear. The simplest way to find out if a new system is suitable is to ask the participants directly. Indeed in certain aspects (e.g., trust and confidence), subjective measures are the only possibility. However subjective data needs to be treated with great caution as it is very easily corrupted by external influences. The experimenter must not simply obtain the opinion; they must also understand where it comes from.

It is essential to ensure that the participants have received sufficient information and training on the test system in order to give an opinion. Most participants will attempt to answer the questions put to them and may not be able to judge whether they have sufficient understanding to do so. They may also be reluctant to be too critical about the system they are asked to evaluate. Experimenter bias can have a big effect here. The more the experimenter is seen to promote the system, the less the participants will be likely to criticize it. First impressions are particularly important: an opinion formed on improper use of a tool may be very difficult to change later on. Moreover, individuals will inevitably be influenced by the comments of their colleagues. In group situations there can be a tendency not to want to oppose the majority view.

bp 30. Minimize the impact of external factors on subjective data

Subjectively expressed opinions are also affected by external factors such as fatigue, platform performance, and mood. Therefore the experimenter should develop sets of probing questions to ensure that the opinions expressed are based on a sound understanding of the system and that there is minimal influence of external factors.

Subjective variables are in general less suitable for repeated measures. If the same questionnaire is presented after each experimental run there will almost certainly be some degradation in the quality of response given. Care must be taken not only in the wording of questions but also in the timing and frequency with which they are asked.

A final point to note in collecting and relying on subjective information. It is clear that air traffic controllers like to solve complex problems (e.g., sequencing, conflict detection and resolution) and do not like the routine tasks such as data block management and data entry. Therefore, any concept that will eliminate or reduce job satisfaction may face resistance. Hence, researchers should be cautious in designing, conducting, and analyzing the results involving such concepts.

5.4 Validation Data Storage

Good data recording and storage are an important feature of any experimental work. In addition to the experimental results obtained, it is essential to keep records of all input data i.e., experimental conditionals, participants, methods, limitations, and any circumstantial information which could influence the interpretation of results. This can be done by maintaining a simple log book, however, the use of electronic formats will make retrieval and dissemination easier.

bp 31. Use a central data repository such as the Validation Data Repository

The collection, retention, and availability of validation data is becoming exceedingly important. The primary benefit of a central data repository, such as the Validation Data Repository (VDR), is to collect this data in one place, where it can be easily searched, retrieved, and analyzed. The use of a central data repository will support:

- Internal project communication – all actors have access to the same data
- Stakeholder communication – for review and dissemination
- Publication – peer group review and publication of results, methods, techniques, tools, scenarios, etc.

The VDR, currently under development, is described in the main body of the document in Section 5.4.

6 Best Practices for After the HITL Exercise

Table 6 presents a list of recommended best practices to be considered when analyzing results and preparing the report for an HITL validation exercise. Each best practice is described in more detail following the table.

Table 6. Best Practices for After the HITL Exercise

6.1	**Managing the HITL Process**
bp32	Report all results, not just those that show an effect
bp 33	Clearly present results to management personnel.
bp 34	Provide results in a timely manner
bp 35	Brief results to participating controllers and pilots.
bp 36	Follow process through to implementation.
6.2	**Statistical and Operational Significance**
bp 37	Ensure that the analyses and results are operationally relevant
bp 38	Only use inferential statistics for pre-planned comparisons
bp 39	Statistical significance doesn't always imply operational significance

6.1 Managing the Process

Regardless of the expected or desired results of a validation activity, it is the obligation of the practitioner to responsibly analyze all data and present all results.

bp 32. Report all results, not just those that show an effect

Even if many dependent variables are used, report all results, not just those that show an effect. Often results of data analyses that show no effect are just as important when operationally interpreted. Regardless, it is good practice to be thorough when reporting the results of data.

bp 33. Clearly present results to management personnel

Results should be presented to management personnel in clear, non-technical terms. In order to convey relevant information, it is important to understand and keep in mind the goals of management. Requesting that management reiterate their interpretation and understanding of the results in their own words will help ensure that there is no misconception of the results and conclusions of the exercise.

bp 34. Provide results in a timely manner

Practitioners recognize the need to provide expeditious analyses to stakeholders and sponsors, but often it is not possible to develop the final report with extensive analyses quickly. Under such circumstances, consider producing two reports, a "quick-look report", and the standard final report. For example, the quick-look report for a full-mission real-time simulation could be primarily based on observations made during the exercise, and the results of questionnaires, interviews and debriefings with the participants. The final report would be expanded to include all sources of data, detailed analyses, detailed results, recommendations, etc. Caution must be exercised in producing a quick-look report. The practitioners and SMEs must carefully consider any interpretations presented since they are based on limited data.

bp 35. Brief results to participating controllers and pilots

After an HITL exercise concludes, practitioners typically do not have any further communications with their study participants. Because data analysis and report generation is often a time-consuming process, this dissociation grows even further by the time practitioners are prepared to release the results.

The fact is, however, that controllers and pilots are usually very interested in knowing the outcomes of the studies in which they participated. They benefit from knowing results because they are essentially spokespeople and information resources on the topic of the evaluation at their respective facilities. In some cases, they may even be able to assist in the training of a new concept or tool because they have had first-hand experience using it in a simulated environment.

HITL practitioners should survey their study participants for interest in obtaining documentation or a briefing at the end of the research and follow through with those requests. Not only will it benefit the controllers and pilots, it will also strengthen the link between the research and operational communities.

bp 36. Follow process through to implementation

If possible, HITL practitioners should track the progress of the concept explored in their exercise to see if and how it gets implemented. They should identify themselves and/or their organization as a resource for information. Decision-makers further down the line are often not aware of or not knowledgeable on the prior testing and may have questions pertaining to the exercise.

6.2 Statistical and Operational Significance of Results

Researchers are expected to indicate the relevance or impact of the results of an HITL exercise in operational terms. In the early stages of an operational concept, data from an HITL exercise is often more descriptive in nature and does not lend itself to abundant statistical analyses. At this stage, results often focus on providing insight into the development of the concept itself and potential operational considerations such as procedural needs, information needs, and impact on human performance. As concepts mature, statistical analyses of data from HITL exercises become more important since quantifiable benefits in terms of safety, capacity, delays, etc., must be also provided by validation exercises. Such analyses require greater scientific rigor.

bp 37. Ensure that the analyses and results are operationally relevant

It is very rare to conduct one simulation exercise to solve complex operational issues or problems. Often a series of exercises are necessary to address some operational decisions. It is critical that the analyses and results of multiple exercises be traced to clear high level validation objectives that relate to operational considerations such as feasibility, safety, benefits assessments, etc. Clear traceability of results from an exercise to relevant operational objectives is necessary.

bp 38. Only use inferential statistics for pre-planned comparisons

In inferential studies, statistical significance helps the experimenter to make comparisons regarding the validity of their experimental hypothesis. Good statistical practice and risk of error dictate that only comparisons planned before the exercise should be considered. If exploratory analyses are used, and interesting comparisons are identified, post developed hypotheses and inferential statistics should not be applied due to the risk of bias and misinterpretation of results. Such valuable information should be used to plan further experimentation.

bp 39. Statistical significance doesn't always imply operational significance

The fact that statistical significance is found does not necessary mean that results translate to meaningful operational differences. For example, consider a case where the workload between two procedures is found to be statistically significant but in both conditions it is very low and would not lead to any operational concerns. The statistical significance will tell you if there is a difference but it will not indicate how meaningful the difference is from the operational perspective. Therefore, statistical significance must be treated with caution.

Often, perception (subjective) responses gathered from questionnaires, interviews, and debriefings help to sort out the operational relevance of observed (objective) results. Therefore, HITL exercises should include subjective data collection strategies designed to substantiate and to interpret the results of the objective tests.

7 Summary of the HITL Exercise Development Process

The development of an HITL exercise is a careful process with many activities that should follow a defined methodology. These activities are described in the context of the High Level Methodology Approach described in the main document:

Step 1. Identify the requirements

- *Define study specific objectives:* A clear and concise statement about the objectives will be developed.
- *Form a team:* The team will consist of members from the operational concept validation management and system analysis team, SMEs, union representatives, researchers, statisticians, human factors engineers, sponsors, and other members. It must be emphasized that all stakeholders must be represented in this team.
- *Identify the type of study:* The team will identify the suitable type of study (e.g., paper study, fast-time simulation, real-time simulation, or rapid prototyping) necessary to accurately assess the operational and technical feasibility of the proposed system changes.

Step 2. Prepare the Validation Plan

- *Develop experiment plan:* An experiment plan detailing the background, objectives, literature review, procedure, data collection and analysis methods, and schedule will be developed.
- *Develop detailed metrics:* The team will identify, define, and develop, as necessary, the metrics required to support the objectives of the study.
- *Develop scenarios and select equipment:* The team will develop air traffic scenarios and select the equipment (e.g., simulator) with due consideration to fidelity requirements.
- *Schedule laboratory and support personnel:* Team will conduct the necessary coordination to ensure that adequate laboratory time is available and support staff will be available when required. This step is typically only required for real-time, HITL simulation studies.
- *Conduct readiness/shakedown testing:* Trial runs of the scenarios will be conducted to ensure that the scenarios, laboratory environment, and operations are realistic. This step

is typically only required for real-time, HITL simulation studies. If necessary, various laboratories need to be integrated and configured to suit study objectives.

Step 3. Carry Out the Validation Plan

- *Conduct simulation and collect data:* Members of the team will conduct the study and collect the data as outlined in the experiment plan.

Step 4. Analyze the Validation Results

- *Analyze data and develop recommendations:* Once the data is collected, members of the team will analyze the data and develop recommendations. Results from multiple studies aimed at evaluating a single operational concept will be reviewed and merged to form a list of recommendations.

Step 5. Prepare the Validation Report

- *Prepare and Publish Reports:* All data, results, and recommendations will be provided to sponsors and other interested parties via published reports.
- *Provide data/information to a central data repository:* Information concerning the exercise will be incorporated into the VDR. Different researchers and organizations will share such VDR information to facilitate validation activities.

8 Overall Summary

The first U.S.-Europe workshop sponsored by FAA/EUROCONTROL Action Plan 5 produced several recommended practices related to HITL exercises. The following topics were discussed in the workshop:

1. Role of HITL exercises in the validation process,
2. General overview of HITL simulation process,
3. Managing HITL exercise,
4. Scope and fidelity considerations,
5. Experimental design considerations,
6. Airspace and scenario characteristics,
7. Subjective and objective data collection,
8. Sources of error and variance, and
9. Statistical and operational significance of results.

Discussion of the first topic captured how HITL exercises could be used for the validation process. The topic on the general overview of the HITL simulation process discussed when, how, and why HITL exercises need to be conducted. The topic on managing HITL exercises provided the practitioners' perspective about how to ensure that HITL exercises address higher-level management questions, researcher questions, and operational questions within the constraints of available resources. The topic on scope and fidelity considerations provided information on how to decide the scope of an exercise and how to ensure and assess adequate fidelity. The topic on experimental design considerations highlighted experimental protocols and considerations particularly applicable to HITL exercises. The topic about airspace and scenario characteristics reviewed the use of generic and real airspace, and highlighted good practices for creating traffic realism and managing scenario characteristics. The topic concerning subjective and objective data collection reviewed the options of when

and how to use the different types data. The topic concerning sources of error and variance identified those areas which should and should not be controlled for in simulation practice. The final topic on statistical and operational significance covered how to report results and interpret operational meaning from statistical information.

From the topics discussed by the practitioners in the workshop, there were many good practices gathered which are reported according whether they apply before, during, or after an HITL exercise.

References

Federal Aviation Administration (1999). *Human Factors Job Aid*.

Federal Aviation Administration (1999). *Operational Concept Validation Process*. FAA-National Airspace (NAS) Advanced Concepts Branch –ACT 540.

Kopardekar, (1999). Simulation Fidelity, Published in Industrial and Occupational Ergonomics: Encyclopedia, ISBN 0-96545-6-0-0.